2011

Measuring Uncertainty of Protein Secondary Structure

Alan Eugene Herner
Wright State University

MEASURING UNCERTAINTY OF

PROTEIN SECONDARY STRUCTURE

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

By

Alan Eugene Herner
B.A. Wright State University, 1980
M.S. Wright State University, 1980
M.S. Wright State University, 2001

2011
Wright State University

WRIGHT STATE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

January 7, 2011

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY ALAN E. HERNER ENTITLED Measuring Uncertainty of Protein Secondary Structure BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

_____
Michael L. Raymer, PhD
Dissertation Director

_____
Arthur A. Goshtasby, PhD
Director, Computer Science and Engineering PhD Program

_____
Andrew Hsu, PhD
Dean
School of Graduate Studies

Committee
Final Examination

_____
Michael L. Raymer, PhD

_____
Gerald Alter, PhD

_____
Travis Doom, PhD

_____
Ruth Pachter, PhD

_____
Mateen Rizki, PhD

ABSTRACT

Herner, Alan E. PhD. Department of Computer Science and Engineering, Wright State University, 2011. *Measuring Uncertainty of Protein Secondary Structure.*

This dissertation develops and demonstrates a method to measure the uncertainty of secondary structure of protein sequences using Shannon's information theory. This method is applied to a newly developed large dataset of chameleon sequences and to several protein hinges culled from the Hinge Atlas. The uncertainty of the central residue in each tripeptide is computed for each amino acid in a sequence using Cuff and Barton's CB513 as the reference set. It is shown that while secondary structure uncertainty is relatively high in chameleon regions [avg = 1.27 bits] it is relatively low in the regions 1-7 residues nearest a chameleon [N terminus flank avg = 1.12 bits; C terminus flank avg = 1.16 bits]. This difference is shown to be highly statistically significant [p = 9.6E-18 and p = 2.9E-12, respectively]. It is also shown that the secondary structure uncertainty of hinge regions was not found to be different to a statistically significant degree once a Bonferroni multiple test correction was applied.

A new hand curated database of long "chameleon" sequences was developed. It contains nine sequences of length eight and eighty-five sequences of length seven.

TABLE OF CONTENTS

Table of Figures

Table of Tables

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Michael Raymer and my mentor Dr. Ruth Pachter for their guidance and patience in showing me the wonders of proteins and bioinformatics. Both spent untold hours teaching me both the content and methods of protein science. I would also like to thank my other committee members Drs. Doom, Rizki and Alter each of whom are extraordinary teachers. It is my hope that someday I may be a teacher who teaches half as well.

I would like to thank the AF Research Laboratory for providing me with not only financial support, but also the time to pursue this goal. In particular, I would like to thank my bosses: Mr. Chuck Wagner, Mr. Scott Pearl, Mr. Brandon Lovett, Mr. Jim Morgan, Mr. Kermit Stearns, and Ms. Persis Elwood who allowed me great flexibility in juggling the demands of work and school.

I would to thank my friends and colleagues at AFRL for their unwavering support of this endeavor. Drs. Gene Himes, John Maguire and Brench Boden all provided help, insight, and encouragement while I worked through this project.

I would like to thank the members of the WSU Bioinformatics Research Group (BIRG) laboratory including Paul Anderson, Gina Cooper, CJ Fravel, Jason Gilder, Amanda Hanes, Ben Kelly, Eric Moyer, David Paoletti, Michael Peterson, Doug Raiford, Sridhar Ramachandran, Deacon Sweeney, and Dan Woldarski. They made it fun.

I would especially like to thank Paul Bender and Dan Schmidt who studied for the qualifiers with me. Their help was invaluable.

I would like to thank my parents who among countless blessings gave me a lifetime love of learning.

Finally, I would like to thank my family for their love and support as I pursued this dream. Thank you Robbie, you are the love of my life. Thank you James, Robin and Thomas, I hope that all of your dreams come true.

MEASURING UNCERTAINTY OF PROTEIN SECONDARY STRUCTURE

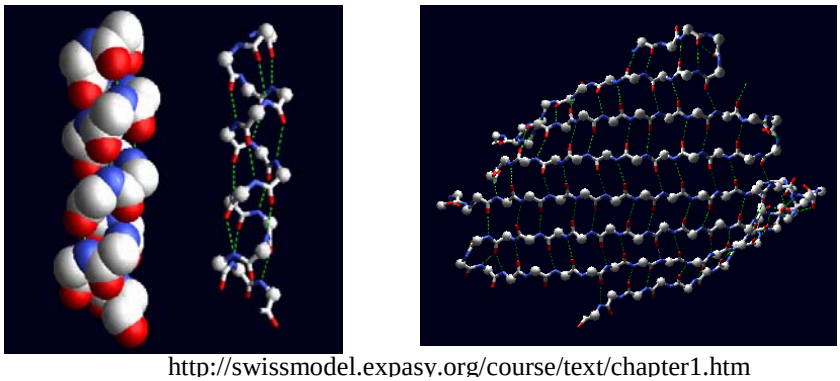
CHAPTER 1
1.0 INTRODUCTION

1.1 Overview

Proteins are one of the essential building blocks of life. Every biological process undertaken by living organisms is in some way mediated, regulated, or facilitated by proteins. As such, one of the fundamental goals of the biological sciences is to gain a better understanding of how proteins are formed, and how they behave. Proteins are synthesized as linear chains (or linear polymers) of amino acids. To achieve a functional state, each chain must first fold into a unique three-dimensional conformation – the protein's native structure. The relationship between the sequence of amino acids which make up a protein chain (often called the protein's primary structure) and its three-dimensional conformation is complex and not well understood. One factor that complicates this relationship is the conformational uncertainty inherent in some amino acid sequences. That is, some sequences have been found to adopt different three-dimensional conformations under different conditions. This dissertation seeks to characterize this *sequence-conformational-uncertainty* using the mathematical formalization of information theory.

The three dimensional conformation of a folded protein can be viewed as a hierarchy composed of several layers. The first layer, as noted above, is the order of the amino acid building blocks that make up the protein chain, or primary structure of the protein. The second layer of protein structure consists of a collection of local conformations that can

be observed in the majority of folded proteins. These elements, termed secondary structure, consist primarily of two folds, the alpha helix and the beta-sheet (Figure 1). There are a variety of software tools available for predicting the secondary structure of a protein based upon the primary structure. These software packages use a wide range of methods from machine learning and pattern recognition to predict which sections of a protein chain will form alpha helices, which will form beta-sheets, and which sections will fold into *coils*, regions adopting neither alpha helical nor beta strand conformation.

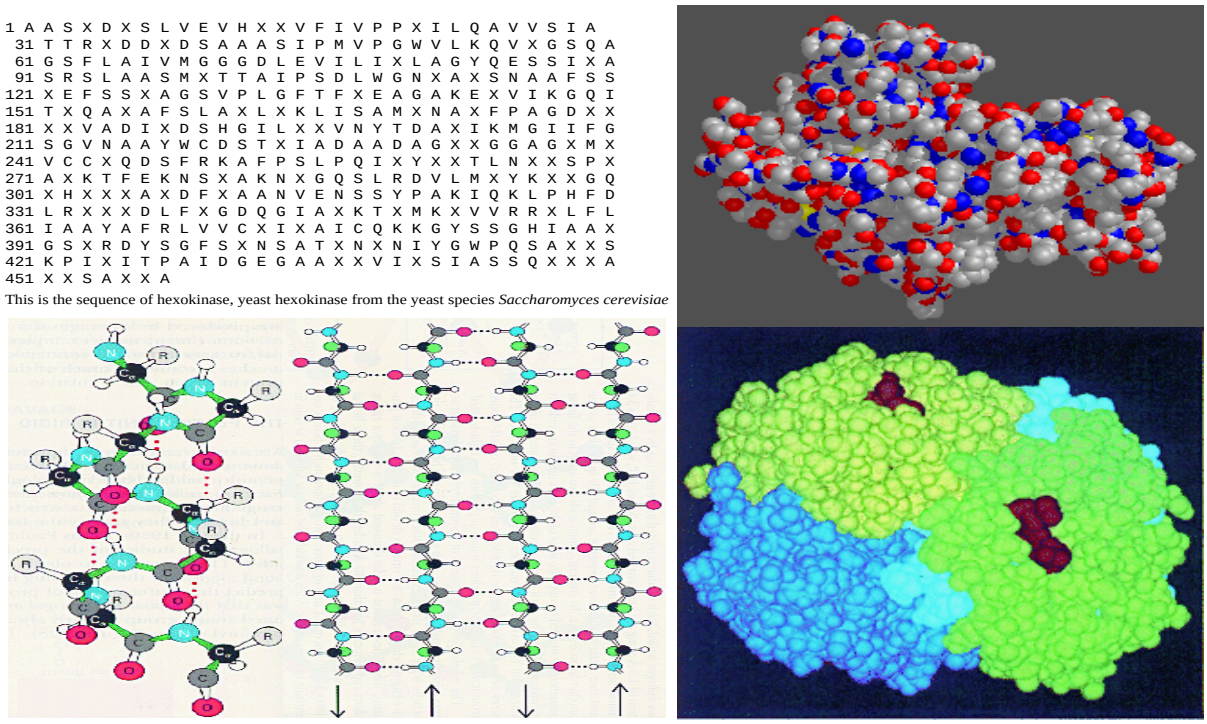


http://swissmodel.expasy.org/course/text/chapter1.htm

Alpha Helix Beta Sheet

Figure 1 - Helix and Sheet

The elements of secondary structure of a protein combine to form the overall three-dimensional shape (or tertiary structure) of an individual protein chain. When more than one protein chain is required to form a functional subunit, the way in which the individual chains associate to form a functional protein complex is called the protein's quaternary structure.



This is the sequence of hexokinase, yeast hexokinase from the yeast species *Saccharomyces cerevisiae*

http://web.archive.org/web/20060411120350/web.mit.edu/esgbio/www/lm/proteins/structure/structure.html

Figure 2 - Protein Structure

1.1.1 Research Objective and Significance

This research was originally motivated by an attempt to understand protein sequences called chameleons. Chameleons are small amino-acid sequences that are known to adopt different secondary structures (helix, sheet, or coil) in different local environments. They are typically five to eight amino acids long. The key question addressed herein is: "How do chameleon sequences compare to "typical" sequences and could I quantify the difference?" This study answers these questions directly, developing a method to measure the uncertainty of protein secondary structure in response to a given primary structure**.** This measure is then employed to characterize two types of protein sequences of particular interest to structural biochemists: 1) chameleon sequences and 2) protein

hinges. Hinges are areas of flexibility which allow two more rigid domains to move relative to one another.

Predicting a protein's secondary structure from its primary structure has become a vital step in investigating the structure and function of proteins. Currently, the state of the art methods are able to achieve around 80% accuracy for this task. The measurement of secondary structure uncertainty may provide important information for comparing protein sequences and identifying critical differences, finding interdomain hinges, and isolating the functional active sites of proteins. This may in turn lead to advances in secondary structure prediction, and in the overall understanding of how proteins find their unique functional conformations.

1.1.2 Organization of the Report

This report is organized into four chapters and six appendices. The first chapter describes basic information about proteins which is important to understanding this work. It includes a short discussion of proteins, protein folding, protein data and databases. Chapter Two is a brief review of the literature associated with secondary structure prediction, Shannon's information theory, chameleon sequences, and protein hinges. Chapter Three depicts the uncertainty measurement method development and use including the development of a chameleon database and the application of the method to chameleons and protein hinges. Chapter Four lists the contributions of this work and describes potential future work. The six appendices (A-F) cover a number of topics related to secondary structure prediction and the formation of proteins.

1.2 Proteins

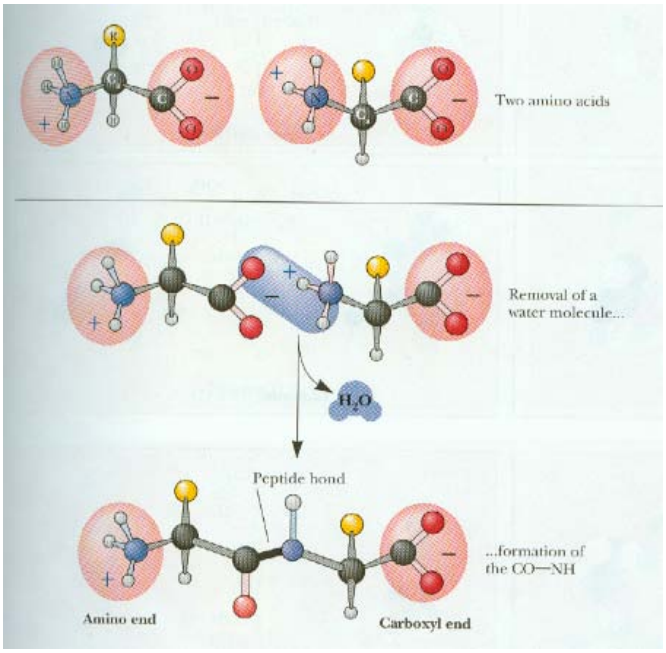


Figure 3 - Amino acid composition and peptide bond
Garret and Grisham, 2005, p.77

1.2.1 Amino acid composition and peptide bonds

There are twenty common amino acids that together make up most proteins (Table 1). Each amino acid consists of an amino terminus (comprising a nitrogen atom and three hydrogen atoms); a carbonyl group; and a central group consisting of a carbon, a side-chain and a hydrogen atom. The central carbon attached to the side-chain is termed the alpha carbon.

As shown in Figure 3, the amino group of a free amino acid has a positive charge while the carboxyl is negatively charged at neutral pH. This allows two amino acids to readily combine to release a water molecule and form a dipeptide in a process known as dehydration synthesis. As a result of dehydration synthesis, the carboxyl group of one of the amino acids is converted to a carbonyl, and a rigid peptide bond is formed between

the two amino acids. Many amino acids can be joined in this manner forming chains (or polypeptides) hundreds of amino acids in length.

One or more such chains, folded into their native conformation(s), then form a functional protein. In globular proteins the native state has several important properties. These include a hydrophobic core, a generally more polar exterior, and in many proteins, one or more catalytic active sites. Finally, the native conformation is stable [Garrett and Grisham 2005 p. 176-8].

Amino Acid	Nonpolar (hydrophobic)	Polar, uncharged	Acidic	Basic	Residue mass (daltons)
Alanine (Ala,A)	X				71.08
Arginine (Arg, R)				X	156.19
Asparagine(Asn,N)		X			114.10
Aspartic acid(Asp,D)			X		115.09
Cysteine (Cys,C)		X			103.14
Glutamic acid (Glu,E)			X		129.12
Glutamine (Gln,Q)		X			128.13
Glycine (Gly,G)		X			57.05
Histidine (His,H)				X	137.14
Isoleucine(Ile,I)	X				113.16
Leucine(Leu,L)	X				113.16
Lysine(Lys,K)				X	128.17
Methionine(Met,M)	X				131.20
Phenylalanine(Phe, F)	X				147.18
Proline (Pro,P)	X				97.12
Serine(Ser, S)		X			87.08
Threonine (Thr,T)		X			101.11
Tryptophan(Trp,W)	X				186.21
Tyrosine(Tyr,Y)		X			163.18
Valine(Val,V)	X				99.13

Table 1 - Amino Acid Properties
Garrett and Grisham 2005, p 78-80
Residue mass Nolting 2006, p.7

1.2.2 Types of Amino Acids

The properties of each amino acid are determined by its side chain. Side chain properties can be categorized along a variety of axes. Among the more common are polar or non-polar, hydrophobic or hydrophilic, charged or uncharged, large or small. Charged amino acids have an overall electrical charge, positive or negative. Polar amino acids are electrostatically neutral overall, but have charged regions or 'poles.' Hydrophilic amino acids easily interact with water via hydrogen bonding. Hydrophobic amino acids do not readily associate with water. Each of these properties has a different effect on the likelihood that a particular amino acid will participate in a given secondary structure.

1.2.3 Planarity and Dihedral angles

The NH group and the C=O group of an amino acid are co-planar with each other and the successive alpha carbon. This allows two degrees of freedom. The angle of rotation around the C_α - N bond is called the psi angle. The angle of rotation about the C_α - C_O bond is called the phi angle. The position of each atom in the main chain can be determined if the positions of a C_α and each of the phi and psi angles are known. This allows for a compact representation for the backbone structure of a protein, as the complete three-dimensional backbone structure can be reconstructed from the phi and psi angles for each alpha carbon. Thus, the conformation of a typical protein chain (three hundred residues) can be approximately represented in only six hundred integer values.

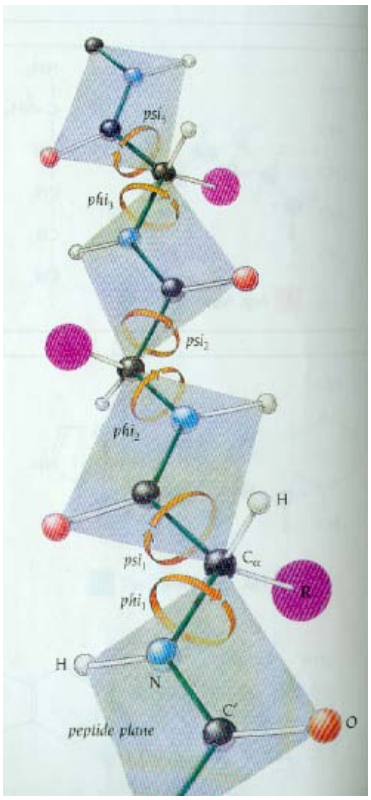


Figure 4 - Peptide Planes and Phi-Psi Angles
Branden and Tooze, 1999, p 8

1.2.4 Conformational Constraints

While phi and psi may independently take any value from -180° to +180° depending on circumstances, the possible combinations of phi and psi are highly constrained. Relatively few combinations are allowed due to steric collisions between the side chains and the main chain of the protein. G.N. Ramachandran was the first person to compute these permissible combinations and the results are illustrated on a Ramachandran plot.

As shown in Figure 5 below, the allowed combinations of psi and phi commonly found in alpha helices, beta sheets and loops or coils are clearly discernable on the plot. Due to its small size, glycine (not shown in figure 5) has a much larger set of allowed combinations

than any other amino acid. This enables glycine to play a unique role in providing flexibility to a protein structure [Branden and Tooze,1999, p 9].

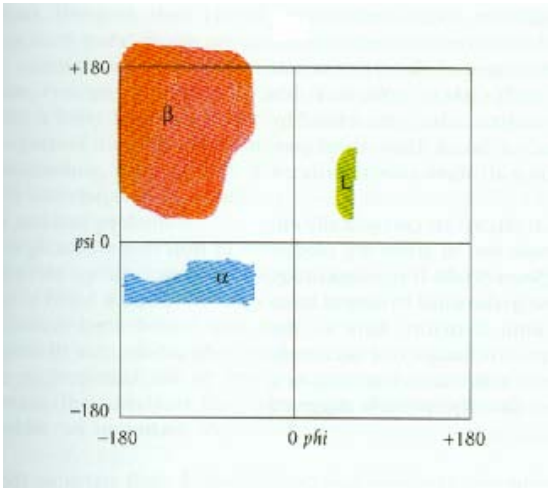


Figure 5 - Ramchandran Plot
Branden and Tooze, 1999, p 9

1.2.5 Molecular forces involved in protein folding

The formation of secondary structure is one step in the overall process of protein folding, in which a linear polypeptide chain folds into its native conformation. In order to develop algorithms to model or predict the outcome of this process, it is necessary to appreciate the physical and chemical forces that drive the process in nature. There are several major forces involved in protein folding. These include hydrogen bonds, the hydrophobic force, ionic interactions (charge), covalent bonds, and van der Waal's forces [Racz, 2007].

1.2.5.1 Hydrogen bonds

Water is a polar molecule. While electrically neutral, a water molecule has a negatively charged oxygen atom, and a region which has a partial positive charge: - the hydrogen atoms.

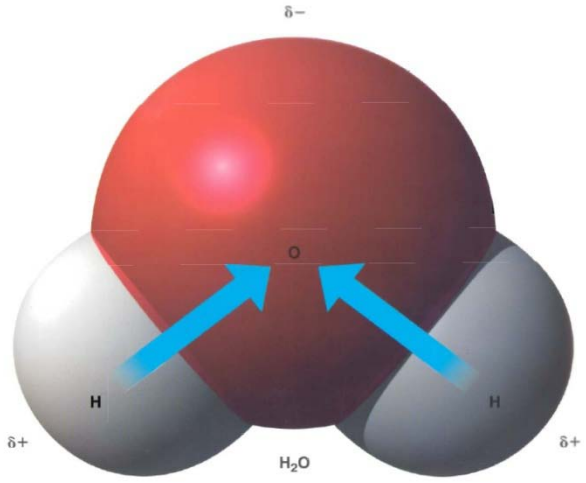


Figure 6 - Polar Regions in a Water Molecule
Campbell *et al.*, 1999, p 29

As a result of this polarity, the hydrogen atoms are attracted to the oxygen atoms of nearby water molecules. This hydrogen-mediated electrical attraction is termed a hydrogen bond. Nitrogen, oxygen and fluorine often participate in polar molecules and hydrogen bonds. Water molecules routinely form hydrogen bonds with other water molecules [Kimball, 2008].

Hydrogen bonds play an important role in stabilizing protein structures. For example, in alpha helices there are hydrogen bonds between the carbonyl oxygen of each of the amino acids four residues further along the peptide chain. Hydrogen bonds among beta strands also serve to stabilize beta sheets.

1.2.5.2 Hydrophobic Effect

A number of amino acids are not polar. These include alanine, isoleucine, leucine, methionine, phenylalanine, proline, and valine [Branden and Tooze, 1999, p 7]. All of these amino acids, with the exception of methionine (S), have only carbon and hydrogen in their side chains. As a result, they do not readily form hydrogen bonds with water. These amino acids are termed "hydrophobic". Sometimes tryptophan is included in this list. [Garrett and Grisham, 2005, p 80].

Faced with a hydrophobic amino acid, a water molecule will 'retreat' to form a hydrogen bond with another water molecule, packing the water more densely. This causes an area to be created around the non-polar molecule which is water-free. This cage-like structure is called a clathrate. Eventually, the water molecules pack as tightly as they can. They then repulse the hydrophobic region, and induce the protein to create a hydrophobic core surrounded by a polar or hydrophilic shell. The hydrophobic effect is generally believed to be an important force in protein folding [Garret and Grisham 2005, p 34-35].

1.2.5.3 Ionic (charge) interactions

Five of the amino acids are charged at biological pH. Aspartic acid, and glutamic acid are negatively charged, while arginine, histidine and lysine are positively charged. When these charges are in close proximity to one another they interact to form salt bridges which stabilize protein structures. The strength of the attraction is governed by Coulomb's law and depends on the inverse of the square of the distance between them. Thus, they are also known as Coulombic forces [Racz, 2007].

$$F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}$$

Where: F = force
q_i = charge on ith particle
r = distance between the two particles
ε_0 = medium permittivity constant

Figure 7 - Coulomb's Law
http://en.wikipedia.org/wiki/Coulomb's_law, Dec 2010

1.2.5.4 Covalent bonds

Among the strongest molecular forces, covalent bonds occur when valence electrons are shared among more than one neighboring atom. Most covalent bonds in proteins are involved in the formation of the protein backbone (N-C_α-C=0). A key exception is the development of disulfide bridges between cysteine amino acids. When the two SH groups at the end of cysteine side chains are oxidated, the hydrogen atoms combine to form water and the sulphurs share electrons. Disulphide bridges often occur in proteins secreted by cells [Branden and Tooze,1999, p 5].

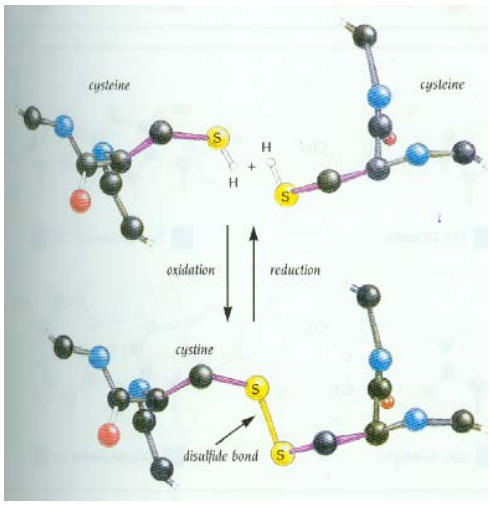


Figure 8 - Disulfide Bridge
Branden and Tooze 1999, p 5

1.2.5.5 Van der Waals forces

Van der Waals forces are weak molecular forces which occur between atoms in close proximity to one another. As parts of a molecule or atom move (*e.g.* an electron), the charges may become unevenly distributed. That is to say, the center of the positive charges in a molecule may not be in the same place as the center of the negative charges, even if they are equal. This causes a temporary dipole to be created which may induce dipoles in neighboring molecules, resulting in an attraction between the two dipole molecules. Van der Waals forces are important in gasses, condensation and liquids.

The native conformations of many proteins have numerous points where amino acids which are sequentially distant are physically in contact with each other. While individually weak, van der Waals forces are often quite numerous and are believed to be a key to stabilizing some protein configurations. Berezovsky and Trifonov have proposed that van der Waals forces may form the basis of a loop-n-lock structure which they use to explain folds in nine different proteins [Berezovsky and Trifonov, 2001].

1.2.6 Protein Structure

1.2.6.1 Primary Structure

The primary structure of a protein is the order of its amino acids in sequence, starting with the N terminus and ending with the C terminus. Anfinsen showed that the information required to achieve a protein's native conformation is encoded in its primary structure (Section 2.1.1.2 Thermodynamic Hypothesis).

1.2.6.2 Secondary Structure

Proteins exhibit several regular patterns which are termed secondary structure. Kabasch and Sander identified eight secondary structures namely, α-helix; β-bridge; extended strand; 3_{10}-helix; π-helix; turn; bend; and other or coil [Kabasch and Sander, 1983, p 2595]. These are usually denoted by their one letter abbreviations: B, E, G, I, T, S, and C, respectively. Given the rarity of most of these structures, compared to the very common alpha helix and extended conformations, these eight structures are often grouped into three more general types: helix, extended strand and random coil (Section 1.3.4 Eight to Three Reduction).

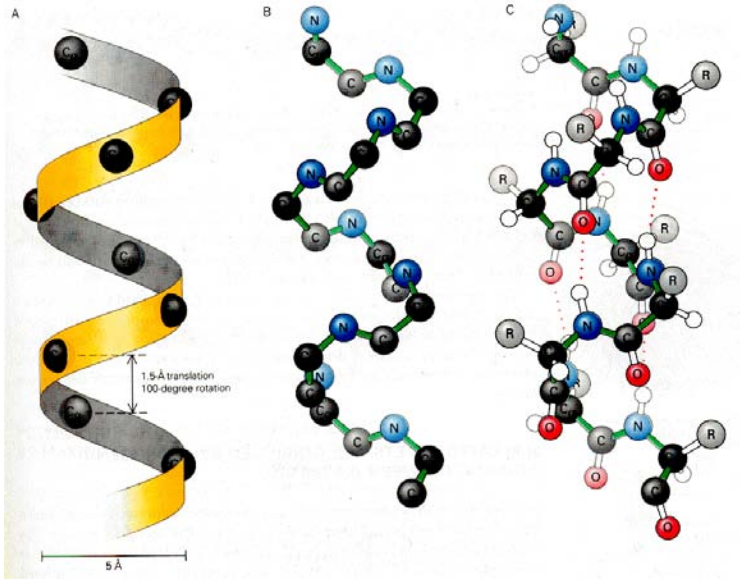


Figure 9 - Alpha Helix
Stryer, 1995, p. 29

1.2.6.2.1 Alpha Helix

Alpha helices are characterized by having 3.6 amino acids per 360 degree turn; phi and psi angles of approximately -60 degrees and hydrogen bonds between the n^{th} amino acid and the $n+4^{th}$ amino acid. The hydrogen bonds are depicted in Figure 9c as dashed red

lines [Stryer, 1995, p. 29]. In addition, they often have hydrophilic (polar) and hydrophobic (non polar) regions (Section 2.1.2.1.1 Helical wheels).

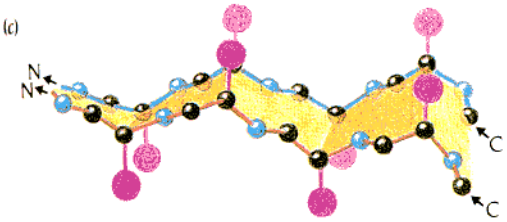
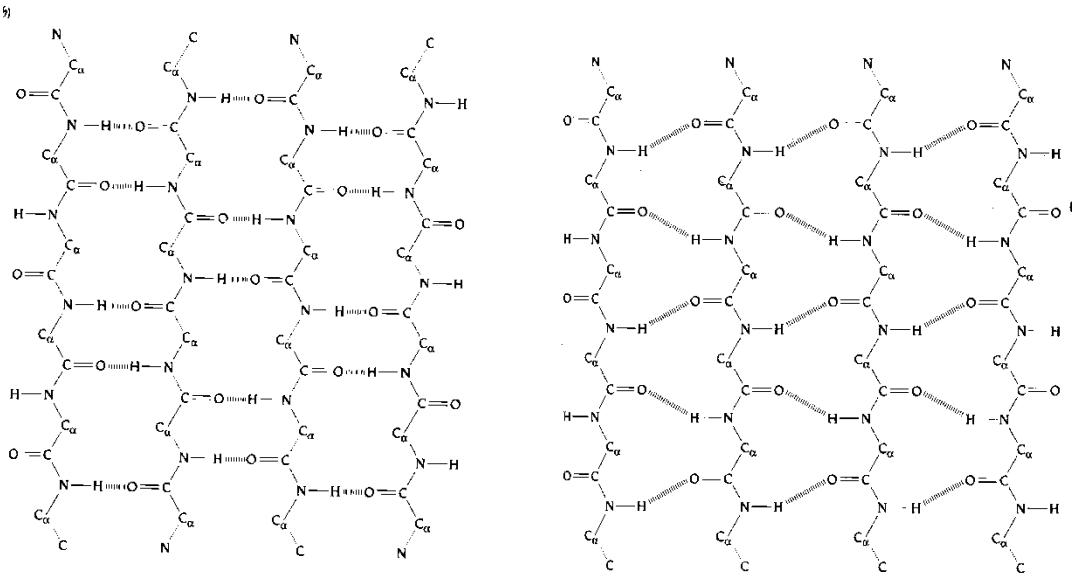


Figure 10 - Two parallel extended strands
Branden and Tooze, 1999, p 19

1.2.6.2.2 Extended strand

Extended strands are generally five to ten amino acids with phi and psi angles of -135 and 135 respectively. They are pleated with one alpha carbon alternating slightly above or below the center line when viewed from the side.



Anti parallel Parallel

Figure 11 - Beta Sheets
Branden and Tooze, 1999, p 18-19

Strands often combine via hydrogen bonds to form β sheets. If the nitrogen ends of the strands are all in the same direction (N-N-N…) the sheet is said to have parallel orientation. If the ends alternate (N-C-N…), the sheet is anti-parallel. The anti-parallel orientation is slightly more stable due to the positions of the hydrogen bonds (Figure 13) [Krane and Raymer, 2003, p 16].

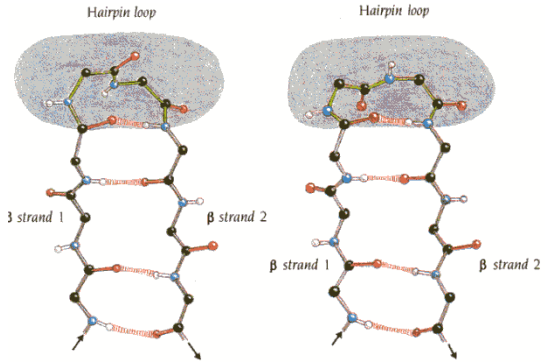


Figure 12 - Hairpin loop
Branden and Tooze, 1999, p 21

1.2.6.2.3 Random Coil

Helices and extended strands are connected to other structures via loops, turns or random coils depending on the length of the connective region. Turns are short, often a few amino acids long. The ends of beta sheets are often connected with structures called hairpin turns or loops. Longer regions are often called loops or random coil. These regions are often highly disordered [Raymer, 2006].

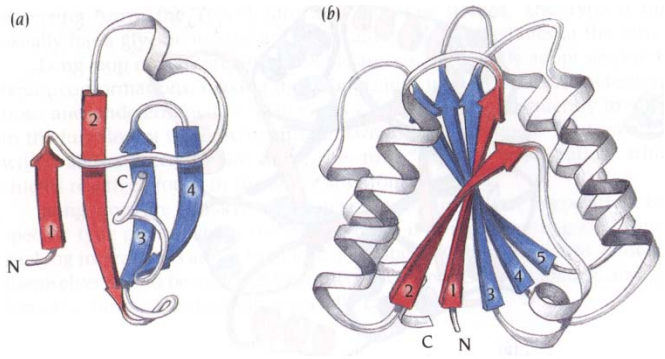


Figure 13 - Antiparallel and parallel beta strands
Branden and Tooze, 1999, p 24

1.2.6.3 Motifs

These secondary structures combine in several different ways to create different groups of structures which occur in many different proteins. Termed motifs, or super secondary structure, these groups have characteristic shapes, structures and names (*e.g*. Barrel, keyhole, Greek key, *etc*.). Some researchers use such motifs as an additional way to categorize and understand proteins. For example, Class, Architecture, Topology and Homologous superfamily (CATH) and Structural Classification of Proteins (SCOP) are two taxonomies that use motifs extensively to classify proteins.

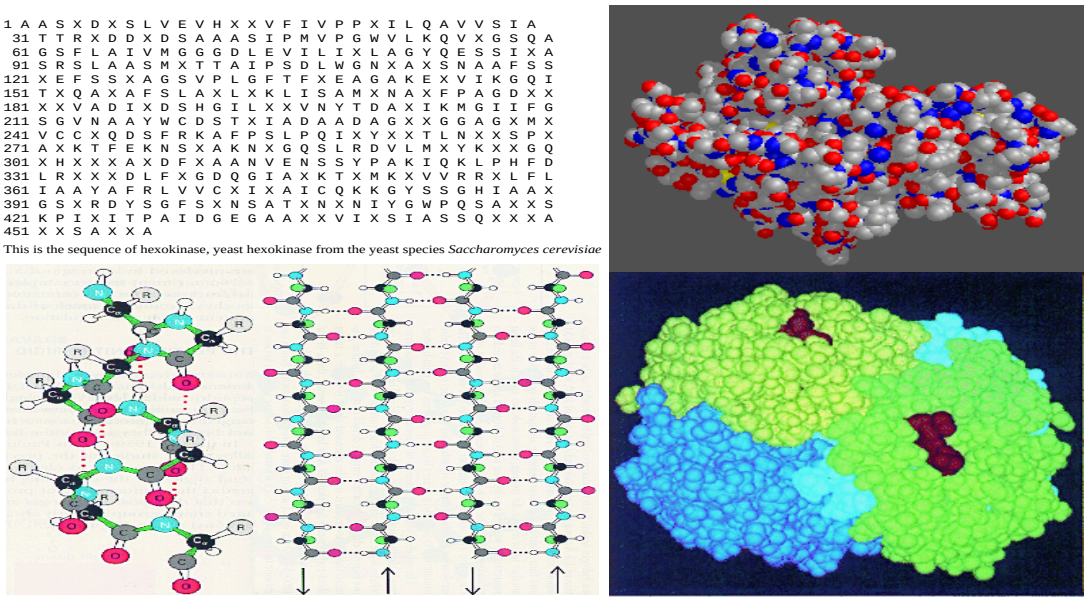
1.2.6.4 Tertiary Structure

Secondary structures combine to form the tertiary, or three dimensional structure of a protein. The tertiary structure determines a protein's function. As a result, the tertiary structure is of keen interest to protein researchers and engineers. It is believed that tertiary structure is defined by the primary structure (Section 2.1.1.2 Thermodynamic Hypothesis). Predicting the tertiary structure directly from the primary structure however, has proven quite difficult. Secondary structure prediction is often a critical step

to successful tertiary structure prediction. For example, Rosetta, one of the state of the art tertiary structure predictors uses secondary predictions of known fragments as a key input (Section 2.12.1.3 Molecular Dynamics).

1.2.6.5 Quaternary Structure

The quaternary structure of a protein is when two or more protein chains are combined together to form a single larger complex or grouping [Krane and Raymer, 2003, p. 14 - 15]. Figure 14 shows examples of each of the four types of protein structure. The upper left shows a typical primary structure. The lower left shows drawings of an alpha-helix and a beta sheet. The upper right illustrates a tertiary structure while the bottom right shows a quaternary structure with the chains color coded.



http://web.archive.org/web/20060411120350/web.mit.edu/esgbio/www/lm/proteins/structure/structure.html

Figure 14 - Protein Structures

1.2.7 Theories of Folding

When investigating secondary structure it is important to understand how proteins fold and what mechanisms are believed to be involved. There are several theories of protein folding. Fersht lists three which have gained prominence in the literature: the framework model; the hydrophobic collapse model; and the nucleation model [Fersht, 1999, p 573-610].

1.2.7.1 Framework Model

Under the framework model there are three steps to protein folding. First, the local secondary structure forms. Then the initial tertiary structure is created, followed by long range interactions which solidify the overall structure. A key notion with the framework model is that the native secondary structure induces the creation of tertiary structure. There are numerous derivative models which improve on various aspects of the framework model.

1.2.7.2 Hydrophobic Collapse Model

In this model, the protein is thought to collapse around its hydrophobic side with the rest of the protein folding up in the space remaining. This theory states that atoms which are distant in the primary sequence but in close physical proximity drive the folding of the protein. Hence, tertiary structure is thought to create secondary structure, not the other way around [Fersht, 1999, p 575-576].

1.2.7.3 Nucleation Model

The nucleation model focuses on the role of local interactions. The idea is that small portions of the secondary structure provide 'seeds' around which the secondary and tertiary structures are grown in a systematic, stepwise fashion. In the classic nucleation model the nuclei are very stable and long range interactions are minimal. In a modification of this model, called the nucleation-condensation model, the nucleus is initiated by neighboring amino acids but is stabilized by longer range interactions. Here, the nucleus and extended structures are stabilized at the same time [Fersht, 1999, p 586].

1.2.7.4 Unified Model

A number of experiments have been done on proteins such as chymotripsin inhibitor 2 (CI2), λ repressor, and barnase which show that each of these models may apply to different proteins, or to the same protein in different circumstances, or at different times in the folding process. With some modification, the hydrophobic collapse and framework models can be made compatible with the nucleation-condensation model and the current experimental data. Fersht lists the following required modifications.

> "…the framework model must be modified so that the formation of secondary structure is linked to the formation of tertiary interactions; and the hydrophobic collapse model must have the formation of tertiary interactions linked to the formation of secondary structure…. Whatever the distinctions of names, stable tertiary and secondary structural interactions must form concurrently."[Fersht, 1999, p 596]

1.3 Protein Data and Databases

When working with protein data there are three additional concepts one should be aware of. These are experimental data, data sets, and eight to three state reduction. Each of these will now be discussed.

1.3.1 Experimental data

Two of the most prominent methods used to measure the positions of atoms in a protein are X-ray crystallography and nuclear magnetic resonance (NMR).

1.3.1.1 X-ray crystallography

X-ray crystallography passes X-rays through a crystal and records the diffraction patterns. These are analyzed to infer the atomic structure of the protein being studied. This is analogous to deducing the structure of an object by taking pictures of its shadow on the wall as it is rotated in a strong light.

X-ray crystallography works well for many types of materials in addition to proteins. Under ideal conditions it can measure positions within a fraction of an angstrom. However, due to their complexity, the best protein measurement resolutions are typically 2-3 Ångstroms.

There are four major steps to X-ray crystallography: 1) growing the crystals; 2) mounting the crystals; 3) creating the diffraction pattern; and 4) analyzing the results. Each of these steps has numerous sub-steps, procedures and precautions. For example, Drenth

[Drenth, 2007, p 2-6] lists four sub-steps to crystallization of proteins: 1) ensure the protein is sufficiently pure; 2) dissolve the protein into a suitable solvent; 3) bring the solution to supersaturation; and 4) grow the crystals. He also lists five separate techniques for achieving these steps: batch, liquid-liquid diffusion, hanging drop, sitting drop and dialysis.

Once the crystal has been formed, it is mounted in an appropriate glass capillary with solvent and an X-ray beam is fired through it. The crystal diffracts the X-ray and the diffraction pattern is analyzed. X-ray crystallography is often a resource and time intensive endeavor. In addition, some proteins have proven to be exceptionally difficult to crystallize. As a result, a computational or experimental alternative to crystallographic methods has been pursued by many researchers in computational molecular biology and biochemistry.

1.3.1.2 Nuclear Magnetic Resonance (NMR)

One alternative to crystallography is nuclear magnetic resonance (NMR). While this method is limited to small proteins (60 kDaltons, and is generally limited to lower resolutions (typically 4 – 6 Ångstroms), it does enjoy a few advantages over crystallography. NMR uses aqueous samples and uses the magnetic resonance of "heavy" isotopes of common elements like hydrogen, nitrogen, or carbon. By incorporating these "heavy" isotopes into the protein and identifying their position in the protein one may infer the structure of the protein. Since the protein is in solution and not crystallized, the folding of some proteins can be inferred. Two weaknesses of the NMR process are that the resolution is not quite as good as X-ray crystallography (4-6

Ångstroms) and it is limited to small proteins [National Institute of General Medical Sciences, 2007 p 26-34].

1.3.2 Dictionary of Secondary Structure of Proteins (DSSP)

The secondary structure (helix, coil, sheet) of a protein is often defined using Kabsch and Sander's DSSP program. The DSSP program is based primarily on hydrogen bonds.

> "Therefore, we base our secondary structure recognition algorithm mainly on H-bonding patterns: "n-turns", with an H-bond between the CO of residue i and the NH of residue i+ n, where n = 3, 4, 5, and "bridges" with H bonds between residues in sequence. … Repeating 4 turns define α-helices, and repeating bridges define β structure, in good agreement with intuitive assignments. All other occurrences of the basic patterns provide and interesting survey of 3_{10} helices, π-helices, single turns, and single β-bridges."[Kabsch and Sander, 1983, p 2578]

In addition, to the patterns defined above, Kabsch and Sander define bends, chirality, and SS bonds based not on hydrogen bonds, but on geometry. For example:

> "Bends are regions of high curvature. We quantify chain curvature at the central residue i of five residues as the angle between the backbone direction of the first three and the last three residues. … For a bend at i, we require a curvature of at least 70°." [Kabsch and Sander, 1983, p 2585]

The DSSP program identifies 8 forms of secondary structure as shown in Table 2.

"H" = 4-helix (α helix)
"B" = residue in isolated β bridge
"E" = extended strand, participates in β-ladder
"G" = 3-helix (3_{10}-helix)
"I" = 5-helix (π-helix)
"T" = H-bonded turn
"S" = bend
"C" = other (not H,B,E,G,I,T,or S)

Table 2 - DSSP Codes

While DSSP is generally accepted as the primary authority for defining secondary structure given the atomic coordinates of a protein, other programs used to define secondary structure include STRIDE and DEFINE. Cuff and Barton compared all three methods on the Rost and Sander RS126 database.

> "When compared pairwise, DSSP and STRIDE agree to 95%, whereas DSSP and DEFINE agree at 73%, with STRIDE and DEFINE agreeing at 73%. All three methods agree at only 71%." [Cuff and Barton, 1999, p 512]

While the other two methods appear occasionally in the literature, DSSP is the oldest and far and away the most commonly used method.

1.3.3 Data sets

Protein secondary structure has all of the normal data concerns which attend an analysis of any kind including: data pedigree, quantity, missing data, errors, outliers, and reconciling data from different sources. Four areas of particular interest are levels of redundancy, degree of homology, data sources, and data formats.

1.3.3.1 Redundancy and Homology

Two issues of great importance to protein modeling are redundancy and homology. When creating or using a database one normally does not want several copies of the same protein or domain. Many copies of the same protein may cause its characteristics to be over represented in the database. Since most models assume that the database on which they are developed is representative of the population of proteins they will used to analyze, over representation of a particular item may bias their results.

Unfortunately, proteins which are homologous (share a common ancestor) are often partially redundant, in the sense that long segments may be the same. This too, can bias a model. While it is not possible to select proteins which are completely non-homologous, care must be taken to select proteins/domains which have a limited degree of homology. The degree of homology/redundancy is usually stated as the percentage of pair-wise positions with matching residues. While strictly speaking this is not correct, it is a general practice. Either proteins share a common ancestor and are homologous, or they do not.

As a result of these concerns, empirical and statistical studies of protein structure are often based upon collections of proteins specifically selected to avoid over representation caused by protein homology. Such data sets are described as being homologous to no greater than some percentage (20, 25, or 40%). This means that given any two proteins in the dataset, no more than a fixed percent of the residues will match when compared position by position.

Some researchers improve this further by performing a multiple alignment analysis using programs such as BLAST or CLUSTAL-W. These programs compare the subject protein residue by residue to a large set of proteins/domains looking for potential homology. By setting appropriate thresholds one can ensure that only a limited degree of homology remains [Jones, 1999].

1.3.3.2 Data Sources

Unlike the protein researchers of twenty or thirty years ago, today's researchers have a plethora of potential data sources. Among these are the Protein Data Bank (PDB), Swiss-Prot, and TrEmble.

1.3.3.2.1 wwProtein Data Bank (PDB)

The three largest protein data repositories, Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (USA), Protein Data Bank Japan (PDBJ), and the European Bioinformatics Institute (EBI) have banded together to create the wwPDB. Recently the Biological Magnetic Resonance Bank (BMRB) (USA) which collects and distributes NMR derived data has joined the wwPDB. As one can see from Tables 4 and 5, the growth in the quantity of protein data has been dramatic over the last fifteen years.

Last Updated: 18 September 2007

Year	Total Depositions	Deposited To			Processed By		
		RCSB	PDBj	EBI	RCSB	PDBj	EBI
2000	2983	2445	10	528	2294	161	528
2001	3286	2673	118	495	2407	384	495
2002	3563	2769	289	505	2401	657	505
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3083	1613	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	6045	3872	1373	800	3575	1670	800
TOTAL	40175	28695	5581	5899	24714	9562	5899

Table 3 - Statistics for PDB Structures
Deposited and Processed By Year and Site
RSCB Annual Report 2008

Year	Total
2000	2631
2001	2835
2002	3013
2003	4182
2004	5213
2005	5405
2006	6542
2007	5052
Total	34873

Note: Experimentally solved structures (excluding obsolete structures)

Table 4 - PDB Structures Released Per Year
RSCB Annual Report 2008

1.3.3.2.2 Customized Data Sets

Unfortunately, while large databases like the PDB are replete with data, they are often unsuitable for use without refinement. Redundancy and homology have been discussed, however researchers have a number of other reasons to develop, filter or refine their data. These include: focusing on specific types of proteins or phenomenon (e.g. transmembrane proteins, chameleon sequences); controlling anomalies (e.g. eliminating short sequences or low information regions); or ensuring a balanced representation among secondary structures.

As a result of these concerns, a large number of custom data sets have been assembled by researchers and groups for their own particular uses. A de facto convention has developed that many of databases are known by their researchers' initials followed by the number of cases in the database. For example, CB513 stands for Cuff and Barton 513. This database has 513 proteins and/or domains. The RS126 database was the collection developed by Rost and Sander to train

their multilevel neural network based classifier Profile network from HeiDelberg (PHD) and has 126 proteins.

These datasets are often combined or filtered to create other datasets with slightly different properties. CB513 was created to control homology. Cuff and Barton started with 1233 domains from the 3Dee database. From these they removed multiple segment domains and those with resolutions greater than 2.5 Ångstroms, leaving 554 (CB554). CB554 was then combined with RS126 and each of the sequences were compared with AMPS. Alignments with a standard deviation score of 5 or greater were eliminated. The remaining sequences were reviewed to eliminate incomplete sequences. They were again compared to RS126. This produced CB513. The 16 domains in CB513 which were less or equal to 30 residues in length were removed to create CB 497.

1.3.3.3 Data Formats

There are several different data formats for protein data. Most are flat files that are easily parsed using Perl or Ruby scripts. Two of the more popular formats are FASTA and Protein Data Bank (PDB).

1.3.3.3.1 FASTA

FASTA is a very simple format. Each entry starts with a line with a less than sign (<) as the first character followed by identifying information. This is followed any number of sequence lines. FASTA is used by many leading analysis programs (BLAST, *etc*.).

1.3.3.3.2 Protein Data Bank

If FASTA represents one end of the format continuum, the Protein Data Bank (PDB) represents the other. The PDB format is very complex. The PDB Contents Guide describing the format is nearly 125 pages long. However, the vast majority of PDB files have four main sections. The first section includes information identifying the protein, who submitted it and references. The second section lists the primary structure as a list of amino acids. The third section lists the secondary structure. The last section contains the atomic data including the amino acids, their atoms, and their position in space. Where a FASTA file may be 5 or 10 lines of data, the corresponding PDB file will run for several pages. Due to the completeness of the data the PDB is normally viewed as authoritative. A number of very sophisticated visualization programs have been developed to make use of information in the PDB file format. One such program is Visual Molecular Dynamics (VMD) from the University of Illinois. Free on the web, VMD reads the atomic data from a PDB file and creates a 3-D visual model of the protein which can be analyzed and manipulated in any number of ways.

1.3.4 Eight to three reduction

As noted previously, Kabasch and Sanders' DSSP algorithm defines eight types of secondary structure. These include H (α helix), G (310 helix), I (π helix), E (extended β strand), B (β bridge), T (turn), S (bend), and C (other or random coil). However, most prediction methods reduce the number of states to three (helix, extended and coil). There are several methods in the literature for accomplishing this reduction.

A: {H,G,I} -> H; {E,B} -> E; {others} -> C
B: {H} -> H; {E} -> E; {others} -> C
C: {H,G} -> H; {E,B} -> E; {others} -> C
D: {H,G,} -> H; {E} -> E; {others} -> C
E: {H} -> H; {E,B} -> E; {others} -> C
F: {H} -> H; {E} -> E; {others} -> C with EE and HHHH -> C
G: {H} -> H; {E} -> E; {others} -> C
with GGGHHHH redefined as HHHHHHH -> C

Table 5 - 8 to 3 Reduction Methods

Of the seven methods listed here the first three (A, B, and C) are far and away the most frequently used.

CHAPTER 2
2.0 LITERATURE REVIEW

There are four areas which touch directly on my work: 1) secondary structure prediction; 2) information theory; 3) chameleons; and 4) protein hinges. The secondary structure prediction literature provides the context for an uncertainty measure. Information theory provides an excellent tool for this purpose. Chameleons and protein hinges are important classes of protein sequences to test the efficacy of the methods developed. Each of these areas will be discussed in turn.

2.1 Secondary Structure Prediction

The literature in secondary structure prediction is both wide and deep. The area is a little over sixty years old and has been widely recognized as a "grand challenge" problem. This review consists of three parts: foundations; illustrative papers representing different approaches and methods; and the current state of the art.

2.1.1 Foundations

2.1.1.1 Early Investigations

The prediction of secondary structure has a long history. One of the earliest investigations was conducted by Pauling and Corey, who in 1951 wrote a set of eight papers on protein structure which were published in the March and May editions of the Proceedings of the National Academy of Science. Three of these are of particular importance to secondary structure prediction.

Pauling, Linus,. Robert B. Corey and H.R. Branson. *"The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain***." Proceedings of the National Academy of Science (PNAS)** (1951) Vol 37 pp 205-211.

Pauling, Linus, and Robert B. Corey. *"Atomic Coordinates and Structural Factors for Two Helical Configurations of Polypeptide Chains***." Proceedings of the National Academy of Science (PNAS)** (1951) Vol 37 pp 235-240.

Pauling, Linus, and Robert B. Corey *"The Pleated Sheet, A Layer Configuration of Polypeptide Chains* **." Proceedings of the National Academy of Science (PNAS)** (1951) Vol 37 pp 205-211.

The first two papers discuss the structure of two types of helices (alpha and gamma) based on their atomic coordinates as measured using X-ray crystallography. The key ideas propounded in these papers was the role of hydrogen bonds in forming and stabilizing helices and the fact that there exists a non integer number of residues in each turn. Pauling and Corey calculated the number of residues per turn for the alpha helix to be 3.7. We now know it to be 3.6. Pauling was given the Nobel Prize in Chemistry in 1954 in part for his efforts in the discovery of the alpha helix.

The third paper discusses the structure of what is now known as the beta sheet. Here too, the key idea was the role of the hydrogen bond in forming and stabilizing the structure of the sheet [Brownlee, 2006].

2.1.1.2 Thermodynamic Hypothesis

Although Pauling and his colleagues had identified a number of the characteristics of protein structure, it remained unclear what mechanisms were responsible for protein structure formation, or folding.

Christian B. Anfinsen and his colleagues investigated this, publishing their work in the September 1961 Proceedings of the National Academy of Science. They found that bovine pancreatic ribonuclease could be reduced with urea and then reoxidized back to its native form.

> "From chemical and physical studies of the reformed enzyme, it may be concluded that the information for the correct pairing of half-cystine residues in disulfide linkage, and for the assumption of the native secondary and tertiary structures, is contained in the amino acid sequence itself."[Anfinsen *et al.*1961, p 1309]

This was followed by a number of studies which culminated in the development of the "Thermodynamic Hypothesis."

> "This hypothesis - states that the three-dimensional structure of a native protein in its normal milieu (solvent, pH, ionic strength, presence of other components such as met al ions or prosthetic groups, temperature, etc.) is the one where the Gibbs free energy of the *whole system* is at its lowest; that is the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in *a given environment."* [Anfinsen 1972, p 104 Italics in the original.]

Some have said that protein chaperones provide a counter example to the thermodynamic hypothesis since they are often required for a protein to achieve its native conformation in nature. However, others argue that chaperones do not create the appropriate environment, they merely maintain it long enough for the required shape to be achieved. Anfinsen was awarded the 1972 Nobel Prize in Chemistry for his work on the thermodynamic hypothesis.

2.1.1.3 Levinthal's Paradox

Cyrus Levinthal did a thought experiment in 1968 which showed that the number of possible angles and hence conformations for a moderately sized protein was far too large

for the protein to do an exhaustive search and select the optimal conformation. In the example he gave the number of possible conformations was 10^{300} and the number which could be reasonably explored in the time observed for folding was 10^{8}. This became known as Levinthal's Paradox. One possible solution to this problem is the idea of an 'energy landscape' where the denatured protein 'falls' into an 'energy well or valley' similar to an object falling down a hill. The intermediate positions may be random and unpredictable, but the end result is the same, a folded protein [Levinthal 1969, p 22-24].

2.1.2 Illustrative Papers

Most methods for predicting secondary structure of proteins can thought of as falling into one of four areas: physico-chemical, statistical, pattern recognition, and ensemble. Papers representing each of these areas are summarized here.

2.1.2.1 Physico-chemical

The three papers here are Schiffer and Edmunson, Lim, and Meiler and Baker. These papers depict helical wheels, physical rules, and molecular dynamics respectively.

2.1.2.1.1 Helical wheels

Building on the work of others, Schiffer and Edmundson developed the notion of a helical wheel in 1967 [Schiffer and Edmundson 1967 p.121-135]. The helical wheel is based on the idea that there are hydrophobic and hydrophilic regions on opposite sides of a helix (a hydrophobic arc). Schiffer and Edmundson recognized that with 100° degrees between amino acids, helices often had hydrophobic residues at positions n, n+/- 3, n+/-4.

This became a relatively easy way to identify helices. A similar idea was developed for beta strands with hydrophobic residues on one side of a sheet (n+2).

2.1.2.1.2 Physical rules

V.I. Lim wrote two classic papers in 1974 [Lim 1974a, Lim 1974b]. In the first paper he states that the Schiffer and Edmundson algorithm is necessary but not sufficient for finding helices. He then identified two main principles governing secondary structure: 1) compactness of form; and 2) presence of a tightly packed hydrophobic core and a polar shell. From these two principles he developed five requirements and approximately 20 complex rules to predict protein secondary structure. (Many of his rules have sub rules, conditions and exceptions.) Lim built a number of physical models ("stick and ball") proteins to develop and test the rules. The rules were built and tested on a database of 25 proteins. Lim claimed 70% accuracy in predicting secondary structure, but researchers using these methods on larger databases report accuracies of approximately 50%.

2.1.2.1.3 Molecular Dynamics

Meiler and Baker describe the use of Rosetta, a molecular dynamics program developed by the Baker laboratory at the University of Washington [Meiler and Baker 2003]. In general, molecular dynamics programs model a protein at the atomic level using Newtonian physics under various conditions (water, temperature *etc*.). Due to the number of possible states to be investigated (Levinthal's paradox) and the number of intermediate steps to be taken for a reasonable level of resolution, molecular dynamics models are typically very computationally intensive and time consuming. There are several ways to reduce the complexity of the problem. One can represent the protein as a

set of rigid bodies, as a collection of centroids, as a set of fragments, or as a set of fully formed structures. Rosetta works by querying a library of protein fragments and building a set of candidate structures obeying various physical constraints. The candidates with the lowest free energy are then selected.

Even with these simplifications, Rosetta is still computationally intensive for any but the simplest problems. As a result, the Baker laboratory has started a Rosetta@home similar to Stanford's Folding@home. In both cases, volunteers install software on their computers that allow researchers to use computing cycles on the volunteers' machines when they are idle. This enables many more problems to be worked than would otherwise be possible.

In this paper, Meiler and Baker use Rosetta to generate tertiary information to help feed a neural network to predict secondary structure. The basic neural network has three stages, 1053 input nodes, 39 hidden nodes and three output nodes. 90 input nodes were added to input the tertiary information. The model was tested on 137 independent proteins.

Using the sequence only, the Q_3 was 75%. Adding data from the Rosetta models the Q_3 was 80%. Using data from the correct native structures, Meiler and Baker report a Q_3 of 82%. One would not normally have the native tertiary structure, but this represents a theoretical maximum for this Rosetta based procedure.

2.1.2.2 Statistical

There are three major sub classes of statistical secondary structure prediction techniques: 1) frequency based (e.g. Chou-Fasman, Bayesian models); 2) information based (e.g. GOR, numerous datamining tools); and 3) linear models (e.g. Discrimination of Secondary structure Classes (DSC), regression). A representative paper in each area is discussed.

2.1.2.2.1 Frequency based

Chou and Fasman wrote two papers in 1974 and one in 1978 which describe a method based on the normalized frequency of each of the twenty amino acids in each form of secondary structure (helix, strand, and coil) [Chou and Fasman 1974a,1974b, 1978]. The normalized frequencies for each amino acid/structure pairing are known as Chou-Fasman numbers. Using these frequencies, Chou-Fasman identified 'nucleating' sites of structure 'formers' which were 'grown' outward until sufficient structure 'breakers' were encountered. The work in 1974 was based on a dataset of 15 proteins. The 1978 paper updated the numbers and the algorithm based on 29 proteins. While Chou and Fasman claimed 75% prediction accuracy, later researchers using larger data bases put the accuracy between 50 and 60%.

In addition, to numbers depicting the propensity of amino acids to participate in helices and sheets the 1978 paper also included numbers for beta turns. Turns are broadly classified by the number of residues participating in the turn. For example: pi turns

involve six residues; alpha turns, five; beta turns, four; gama turns, three and delta turns involve two residues.

The Chou Fasman numbers and their notions of structure formers and breakers have been a classic way to characterize protein sequences through the years. These were used to compare to the information metric developed as the thrust of this work.

A.A.	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.07	0.106	0.099	0.085
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.07
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.12	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

http://course.wilkes.edu/bioinformatics/stories/storyReader$122

Table 6 - Chou Fasman Parameters (1978)
29 Proteins

2.1.2.2.2 Information theory

The Garnier, Osguthorpre and Robson (GOR) algorithm is an important prediction model based on information theory. For completeness it is described here. Information theory will be presented in more detail in Section 2.2.

The basic idea behind the GOR model is that conformation of a given a residue in a protein chain may be predicted using information from other residues in the same chain. What is different about the GOR method is it uses formal information theory to quantify the information gained by additional residues to the mix.

> "The most general statement concerning the information we have for the conformation of the jth residue is thus
>
> $$I(S_j: R_1, R_2, \ldots R_{last}),$$
>
> which reads as the 'information which the first, second, and so on up to the last position carry about the conformation of the jth residue [Garnier *et al.* 1978, p 99]."

Based on earlier research by the one of the authors it was shown that most of the information about the conformation of a particular residue j was to be found in the eight residues on either side of j. Hence, the original GOR algorithm looks at a 17 amino acid window to predict the middle residue[Garnier *et al.* 1978].

> "…the information function I(S,R): I(S;R) = log [P(S|R)/P(S)]." [Kloczkowski *et al.*, 2002, p 157]

This formula is then used to calculate the information contributions of each of the amino acids in the 17 amino acid window to the conformation state of the middle amino acid. Tables are presented with the relative contribution of each of the 20 naturally occurring

amino acids to each of the possible conformations of j given their distance from j. The original GOR paper predicted helix, sheet, coil and reverse turn. Reverse turn is often classified as coil. Each of the tables is then examined and the conformation with the highest information content is the prediction.

The current version of GOR (GOR V) includes five key improvements:

1) using the Cuff and Barton database of 513 non redundant domains;
2) optimizing the decision parameters for the new data;
3) including triplet statistics;
4) defining a resizable window; and
5) using the results of PSI-BLAST multiple alignments with the GOR algorithm.

As a result of these improvements GORV enjoys a Q_3 of 73.5 % [Kloczkowski *et al.* 2002, p 159-161].

2.1.2.2.3 Linear models

Qin, He and Pan developed a two stage linear regression model to predict secondary structure [Qin *et al.* 2005]. They start with multiple alignment data from PSI-BLAST and add chemical properties such as hydrophobicity and mass. Given the size of the window (17), the number of amino acids (20) and the number of potential interaction terms (272) the data is then used in 3 very large (612 coefficients) multiple regression equations. There is one equation for each type of secondary structure, helix, strand, and coil. The data for each amino acid is entered into the three equations and the one which results in the highest value is the first stage prediction.

The results of the first stage prediction are then fed, with the original data, into a second set of multiple linear regression equations. The output of this second set is compared to

generate the predictions of the system. Qin *et al.* validate the system using the leave one out method. They report a Q_3 of 76.4 % and a SOV of 73.2%

2.1.2.3 Pattern Recognition

The pattern recognition area is quite broad. It includes methods such as k-nearest neighbor (K-NN), artificial neural networks (ANN), and hidden Markov models (HMM). Of these, the neural networks using multiple alignment data have achieved the highest Q_3 scores.

2.1.2.3.1 K-Nearest Neighbor

Yi and Lander developed a nearest-neighbor algorithm to predict secondary structure. It used a database of 110 protein chains. Each amino acid was depicted by a three dimensional vector of secondary structure, accessibility, and polarity. The K-nearest neighbors were identified based on a scoring table and the appropriate label assigned. Yi and Lander achieved a Q_3 of 68%. Nearly 4% better than its best predecessor, it was immediately eclipsed by PHD, which was published in the same volume of the same journal [Yi and Lander 1993].

2.1.2.2.2 Neural Networks

Most of the state of the art prediction programs are currently neural nets. Rost identified ninety-nine (99) papers applying neural networks to protein prediction or classification published between 1988 and 2001. Roughly one quarter of these are predicting secondary structure [Rost 2003].

The first program to achieve a Q_3 of 70 %, Profile network from HeiDelberg (PHD), was unparalleled in secondary structure prediction (Q_3) for half a decade. Rost and Sander did this by incorporating evolutionary information through multiple alignment data into a three level neural network. They also used a carefully screened database of 126 non-homologous proteins [Rost and Sander, 1993]. Even today the newest iteration of PHD, called PROFphd, is among the leaders.

2.1.2.2.3 Hidden Markov models

The Baker laboratory also developed a hidden Markov model, HMMSTR, as a method to model tertiary protein structure. They begin with a library of structure invariants or I-sites. These are converted to a series of Markov chains which are merged when overlap is found. Each state has four variables: amino acid, secondary structure, backbone angles (Phi-Psi), and structural context (middle vs. end of strand *etc*.). Each variable has a particular probability distribution and corresponds to a separate model.

The Markov models are then run and the relevant predictions are made. Bystroff *et al.* list six HMMSTR applications: gene finding, secondary structure prediction, structural context, dihedral angle region prediction, protein design, and sequence comparison. As a secondary structure classifier HMMSTR is quite successful, achieving a Q_3 of 74.3% [Bystroff *et al.*, 2000].

2.1.2.2.4 Ensemble Models

Ensemble or composite models combine the results of several other models to develop a 'consensus' prediction through a voting scheme. While voting is a fundamental part of

many algorithms, ensemble models are limited to those methods where the underlying models are stand-alone and external to the method itself. Ensemble models work best when the underlying programs are orthogonal. Ideally, the input models would make errors which were different, complementary and predictable. To the extent that this can be accomplished, ensemble models can greatly improve the resulting predictions.

Cuff and Barton's JPRED is the classic example of an ensemble model. It works by combining the results of four other classifiers (PHD, DSC, PREDATOR, and NNSSP) in a simple majority wins arrangement. This results in a Q_3 of 72.9%. Adding ZPRED, refining the input data and voting methods, and using JPRED predictions themselves as inputs, Cuff and Barton increased the Q_3 achieved in later versions (JNET) to 76.4% [Cuff and Barton, 1999].

2.1.3 Current State of the Art

2.1.3.1 Q_3 77% - 81%

The current state of the art is a Q_3 of 77-80%. There are twelve efforts that have achieved a Q_3 of 77% or better. Two of them, PROF and EVA were developed by Rost and his team. Two others, SSpro and Porter were developed by Pollastri *et al.* Three are networks of binary classifiers (H, ~H) using support vector machines (Hui *et al.*, Wang *et al*., and Kim and Park). The others have been developed by Petersen *et al*.; Jones (PSIPRED); Montgomerie *et al.* (PROTEUS); Dor and Zhou (SPINE); and Wood and Hirst (DESTRUCT). Two, EVA and PROTEUS, are ensemble models, the others are standalone pattern recognition methods (multi-level neural networks or support vector machines).

2.1.3.2 PSIPRED

PHD uses BLAST alignments to get the evolutionary information into a multistage neural net. This resulted in a six percentage point improvement in Q_3 over other methods. Jones was the first to recognize that an intermediate product of the BLAST software, the PSI-BLAST log file, held even more useful information.

The PSI-BLAST log file is a position specific scoring matrix (PSSM). It contains the log-likelihood of each amino acid to be found in each position. This enables information from a large number of non-redundant proteins to be applied to create the secondary structure prediction for each residue. PSI-BLAST's iterated process incorporates more distant homologues.

PSIPRED filters out low information regions and transmembrane proteins. It then feeds the data into BLAST and outputs the PSSM. This is input into a multi-stage neural network. The result is that PSIPRED achieves a Q_3 score of 76.5 when the H,G -> H; E,B -> E; others -> C reduction is used; and 78.3 when the H-> H; E-> E; others -> C reduction is used. PSIPRED also did very well at the third Critical Assessment of Techniques for Protein Structure Prediction (CASP3). CASP is a biannual competition of protein classifiers [Jones, 1999].

2.1.3.3 PROFphd

PROFphd is an updated version of PHD. Quali and King developed a very good secondary prediction model at approximately the same time that is also called Prof. Rost calls his model PROF or PROFphd and the Quali and King model Prof King. Among the

improvements Rost implemented in PROF was to use the PSI-BLAST profiles. This resulted in a Q_3 of 77% [Rost, 2003].

2.1.3.4 EVA-4

EVA is a large web based tool set which automatically collects all of the new additions to the PDB and runs many of the web based protein classifiers against them. Rost ran four of the best secondary structure classifiers (PSIPRED, PHDPSI, SSPRO, and PROF) against the EVA data and averaged them. It produced a Q_3 of 77.8% [Rost, 2003].

2.1.3.5 SSpro

SSpro, uses a collection of eleven bi-directional recurrent neural networks (BRNN). In a BRNN the output for t + 1 case and the t -1 case are also inputted into the neural net along with the data for case t to determine the output for case t. The forward (t+1) and backward (t-1) contexts are both implemented as forward propagation networks. Each of the BRNNs has a different number of hidden units, output units and weights.
Baldi *et al.* built their own training set using 1180 sequences from the PDB (1999) which were:

> " …(a) at least 30 amino acids long, (b) have no chain breaks(defined as neighboring amino acids in the sequence having a C^{α} -distances exceeding 4.0 Å), (c) produce a DSSP output, and (d) are obtained by X-ray diffraction methods with a resolution of at least 2.5 Å. Internal homology was reduced by using an all-against-all alignment approach, keeping the PDB sequences with the best resolution. A 50% threshold curve was used for homology reduction. Furthermore, the proteins in the set have < 25% identity with the sequences in the set R126."[Pollastri *et al.*, 2002, p 229]

In addition to using the 8 to 3 reduction H,G -> H; E,B -> E, all others -> C; SSpro also has a version which predicts all eight of the DSSP structures (SSpro 2.0). Pollastri *et al.*

report a Q_8 of 62.58 on the RS126 data set. They report a Q_3 of 78.13 on the same data set [Pollastri 2002].

2.1.3.6 Porter

Porter is a direct descendent of SSpro 2.0. Improvements over SSpro include the following:

1) a near doubling of the size of the database from 1180 to 2171 proteins;
2) expansion of the initial amino acid alphabet to include B,U,X,Z and .(gap);
3) replicating the five two stage bi-directional neural networks (BRNN) nine times each to create an ensemble of 45 BRNNs; and
4) averaging the results of multiple windows as a filter.

Porter uses the more difficult reduction of H, G, I -> H; E,B -> E and others to C. This is the same reduction which was used in the CASP competition when CASP ran secondary structure prediction competitions. Porter uses a PSSM as input to the BRNNs. Pollastri and McLysaught use a five-fold cross validation to test Porter.

These improvements result in an overall Q_3 of 79.01%. Pollastri and McLysaught state that using Petersen's reduction of H->H; E ->E; others -> C allowed them to surpass a Q_3 of 81% [Pollastri and McLysaught, 2005].

2.1.3.7 Petersen *et al.*

Peterson *et al.* developed an ensemble of feed-forward two stage neural networks. Each stage has one input layer, hidden layer and one output layer. Each of the neural networks has a different window size, (15,17,19,21), a different number of hidden nodes (50 or 75) and the same 9 output nodes corresponding to the three possible states of the three central amino acids (i-1,i,i+1). The output nodes of the first stage become the input nodes for the

second stage. The second stage consists of the 9 input nodes, a hidden layer of 40 nodes and an output layer of three consisting of H, E, or C.

The eight networks described above were trained on a homegrown set of 1032 protein chains. The 1032 proteins were divided into ten groups. The networks were then trained and tested in a ten fold cross validation process. Peterson *et al.* used the same process for both the first and second stages. This resulted in a total of 800 predictions for each position. These predictions were used to compute a probability matrix which was used to predict a new test sequence (RS126). Output expansion, *i.e.* predicting not only the ith amino acid but also the ith–1 and the ith+1 amino acids, was also used.

Peterson *et al.* used a voting scheme among the 800 predictions which computed the reliability of the prediction (ie. highest probability – next highest probability). The average and standard deviation for all predictions in a chain are calculated. If the reliability of the prediction is greater than the mean plus one standard deviation, it is added to a weighted average for the position being predicted. This results in very confident predictions being given much greater weight.

Peterson *et al.* report that, using the H-> H; E -> E and others to C reduction assignment, their method achieved a Q_3 of 80.2% on the RS126 data set. Using the reduction of H,G -> H ; E,B -> E and others to C, they achieved a per residue Q_3 of 77.2% [Petersen *et al.* 2000].

2.1.3.8 PROTEUS

Published in 2006, PROTEUS is an ensemble method which combines the results of PSIPRED, JNET and TRANSSEC. PSIPRED is among the best standalone models. JNET is itself a ensemble model combining PHD, DSC, PREDATOR and NSSP, and TRANSECC is a three stage neural network which they developed in-house. Proteus incorporates two key improvements, namely a jury of experts program and a homology search program. These will now be discussed.

The jury of experts (JOE) program is a simple feed forward neural network which combines the results of the three input systems using a single hidden layer. The homology search routine attempts to exploit any similarity between the input sequence and known structures. If a similarity is found, the known structure is substituted for the prediction. This is unique. Since homologous proteins are known to often share similar secondary structure, most researchers attempt to identify and eliminate homologous proteins upfront, using the known structure to predict the sequence structure directly. While theoretically, organisms either share a common ancestor or they do not, as a practical matter homology is often a matter of degree or evolutionary distance. As a result, using limited homology to improve a prediction over a short range of amino acids may have value.

PROTEUS takes three good classifiers PSIPRED, JNET and TRANSSEC and makes them better by combining them and exploiting homology. Montgomerie *et al*. ran four sets of tests on the PROTEUS system. On a test set of 125 randomly selected sequences they report Q_3s for PSIPRED, JNET and TRANSSEC as 78.1, 73.2 and 70.3%

respectively. When combining these predictions without homology, PROTEUS has a Q_3 of 79.7%. When the homology search routine is employed, a score of 87.8% is achieved. Surveying the results of all four test sets, Montgomerie *et al.* state:

> "When restricted to sequence-unique proteins (such as those found in EVA or those targets selected for structural genomics projects) PROTEUS has a Q_3 of 81.3%, which is 4-8% better than the best performing methods. When allowed to predict the structure of any generic protein (as might be done for a genomic annotation project) PROTEUS has a Q_3 of 88-90% which is 12-15% better than the best performing methods described to date."[Montgomerie *et al.* 2006]

2.1.3.9 SMVpsi

Kim and Park [2003] developed a set of binary classifiers based on support vector machines (H/~H, E/~E, C/~C, H/E, H/C and E/C). They used RS126 and CB513. In addition, they developed two new data sets, KP480 a subset of CB513 and a new set of 136 sequences which they used in a blind test. Kim and Park used the PSSM data generated by PSI-BLAST. They combined the binary classifiers into nine different configurations and tested each one. Several window lengths were tested. Jury voting among the classifiers was employed. Seven fold validation was used.

They trained and tested on RS126, CB513, and KP480 achieving a Q_3 of 76.1, 76.6 and 78.5 respectively. Unfortunately, the 8 to 3 reductions were not the same and therefore the numbers are not directly comparable. For RS126 and CB513 H,G -> H; E,B -> E; others to C. For KP480 H -> H; E,B -> E; all other states to C was used. Nevertheless, they are competitive when compared to others in the literature on these same data sets.

2.1.3.10 Wang *et al.*

Wang *et al.* built on the work of Kim and Park. Wang *et al.* also developed a collection of binary classifiers (H/~H, E/~E, C/~C, H/E, H/C and E/C) based on support vector machines. They combined them in six different configurations and compared the results. They used the RS126 and CB513 databases to train and test on. They used a radial basis function as a kernel in the SVMs. The reduction was H, G -> H; E, B -> E; others -> C. Several windows were tested and the best lengths varied from 11 to 19 depending on the classifier and configuration used. The model was validated using a seven fold cross validation.

One of the things which make Wang *et al.*'s method particularly noteworthy is unit of analysis. Wang *et al.* use the amino acid as the unit of analysis, computing a normalized probability of helix, sheet and coil in much the same way that Chou-Fasman numbers are calculated for each of the databases. To this they add the Kyte-Doolittle hydrophobicity numbers. Unlike Kim and Park, and most modern analyses, they do not use PSSM numbers. Yet with a Q_3 of 78.44% they are highly competitive [Wang *et al.*, 2004].

2.1.3.11 DESTRUCT

Wood and Hirst [2005] have developed DESTRUCT (Dihedral Enhanced STRUCTure prediction). DESTRUCT predicts the psi angles of a protein and then uses the prediction to predict secondary structure. It is a multilevel cascade-correlation neural network. A cascade-correlation neural network differs from a back- propagation network in a number of ways. Chief among them is the fact that a back-propagation has a fixed topology and variable weights where as a cascade-correlation has fixed weights and a variable

topology. That is to say, nodes with fixed weights are added to the network to improve performance.

DESTRUCT uses a group of eight nodes as pool of candidates for incorporation. The best node and its weight is added the network. The performance of the network is evaluated and additional nodes are added until an accuracy threshold is met or a maximum hidden node count is reached.

CB513 data is used to train the model. No reduction method is identified. However, CASP 4 and 5 data is used test the method. The reduction used for CASP is H, I, G -> H; E,B -> E. A modified 10 fold cross validation was used to validate the results. A tenth was used for validation, a tenth for selecting the new nodes, and four-fifths for training the network.

DESTRUCT uses the PSSM to predict the psi angle associated with each amino acid and a first estimate of secondary structure (helix, sheet, or coil). A window of 15 is used. The results are then filtered to smooth the prediction using 10 rules based on a window of 7. The psi angle and secondary structure predictions are then combined again with the PSSM data to predict the psi angles a second time. These predictions are iterated four times resulting in a Q_3 of 79.4%.

2.1.3.12 SPINE

SPINE(prediction of Structural Properties of proteins by Integrated NEural networks) is a system of two two-stage neural networks. Both neural networks are back-propagation

with sigmoid activation functions. All initial input and output values were generated using a random number generator.

To train SPINE Dor and Zhou used a non-redundant set of 2640 proteins with less than 25% similarity. The sequences were run through BLAST to generate PSSM data. This was combined with data on seven amino acid properties. These properties included a steric parameter, hydrophobicity, volume, polarizability, isoelctric poin, helix probability and sheet probability. The PSSM and properties data was then used to feed the two sets of neural networks.

The 8 to 3 reduction used was G,H,I -> H; E,B -> E, Others -> C. Dor and Zhou randomly selected 5% of the data for testing from each training session. Weights which were successful in predicting the 5% were saved for use later. Several window sizes were tested. Both 100 and 200 hidden units were tested at the first level, with two hundred proving to be slightly better under all conditions tested.

SPINE was tested using tenfold cross validation. Against their 2640 protein database Dor and Zhou report a Q_3 of 79.5%. For proteins of moderate size (50-300 amino acids) they report a Q_3 of 80.0%. They also report Q_3's of 77.07% and 76.77% on Carugo-338 and CB513 respectively [Dor and Zhou, 2007].

2.1.4 Secondary Structure Prediction Literature Review Summary

Starting with Pauling and Corey's discovery of alpha helices in the fifties numerous researchers have attempted to predict the secondary structure of a protein from its

primary structure. There are literally hundreds of papers on the subject. In this review, four broad methods have been identified to classify the efforts in this area. These are: physico-chemical, statistical, pattern recognition, and ensemble methods. Representative papers in each area were discussed. Finally, twelve papers which form the current state of the art (Q_3 of 77+%) were discussed in detail.

From this review the following conclusions may be drawn. The most successful methods currently use data which has limited internal homology, usually < 25%. A common data set such as CB513 is often used for training or testing. Recent efforts have also used newly deposited proteins from the PBD. The eight to three reduction assignment methods most frequently used are: H,G -> H; E,B -> E; others to C and the plain reduction H->H; E -> E; others to C.

Rost and Sander demonstrated the superiority of using multiple alignment data in training ones classifiers in PHD. Jones showed the power of using the PSSM data generated from PSI-BLAST. Each of the state of the art methods now uses this data. They also use some form of neural network or ensemble methods which employ neural networks as root methods. The transformations used vary from method to method. Some use reliability information, some use probability matrices, others various voting methods. Most of the state of the art methods are validated using a n-fold cross validation where n is 7, 10 or leave one out. Some methods use a separate test set. RS126 is popular but may be less demanding than other test sets. All of the state of the art methods are less than transparent. This is due to the fact that they are all, at some level, neural networks. This

causes them to be black boxes. Lastly, all of the state of the art efforts enjoy Q_3 scores of 77% or greater.

		Illustrative Secondary Structure Prediction Efforts							
Method	Program	Author Year	Data	Reduction	Unit of Analysis	Transformation(s)	Validation	Transparency	Accuracy Q3 (by residue)
Physico-Chemical Rules	Helical Wheels	Schiffer and Edmundson 1967	7 proteins	none	single amino acid	window of i+/- 3 or 4	7 proteins	High	NA
Physico-Chemical Rules	Lim Rules	Lim 1974	25 proteins	none	single amino acid	numerous hand crafted rules	25 proteins	High	~50%
Physico-Chemical Molecular Dynamics	Rosetta	Meiler and Baker 2003	~1000 proteins	none	multiple alignment PSSM	window of 39 tertiary information from Rosetta is added to seven amino acid properties and put into a three stage neural network	137 proteins	Moderate	75% sequence only 80% sequence plus rosetta models 81% sequence plus native structures
Statistical Frequency Based	Chou-Fasman	Chou and Fasman 1974/1978	15/29 proteins	none	single amino acid	numerous rules	15/29 proteins	High	50-60%
Statistical Information Based	GOR V	Kloczkowski etal. 2002	CB513	H->H E->E strings of H less than 5 or E less than 3 -> C others -> C	multiple alignment PSSM	window of 17 amino acids; information computation	leave-one-out cross validation	Moderate	73.5%
Statistical Linear Models	Multiple Regression	Qin etal. 2005	CB513	H,G,I -> H E,B -> E others -> C	multiple alignment PSSM	window of 17 amino acids; 3 very long equations (612 coefficients)	leave-one-out cross validation	Moderate - High	76.4%
Patttern Recognition K-Nearest Neighbor	K-Nearest Neighbor	Yi and Lander 1993	110 proteins	H,G,I -> H E,B -> E others -> C	amino acid sequence;	two scoring systems; 3 window sizes; 3D environment variables; Combined 6 KKN predictions with neural net	110 Protreins	Moderate	68.0%
Patttern Recognition Neural Network	PHD	Rost and Sander 1993	RS126	H,G,I -> H E,B -> E others -> C	multiple alignment	window of 13	7 fold cross validation	Low	70.0%
Patttern Recognition Hidden Markov Model	HMMSTR	Bystroff etal.	618 proteins	none	3D structural motifs	3 D pattern matching against a library of motifs	61 randomly selected proteins (10%)	Moderate	74.3%
Ensemble Model	JPRED/JNET	Cuff and Barton 1999/2000	CB513/CB480	H,G -> H E,B -> E others -> C	multiple alignment PSSM	JPRED combines results from PHD,DSC, PREDATOR and NNSSP JNET adds ZPRED to results of JPRED	CB480 for 7 fold cross validation and CB406 for independent validation	Low	76.4%

Table 7 - Illustrative Structure Prediction Efforts

		State of the Art Secondary Structure Prediction Programs							
Method	Program	Author Year	Data	Reduction	Unit of Analysis	Transformation	Validation	Transparency	Accuracy Q3 (by residue)
Patttern Recognition Neural Network	PSIPRED	Jones 1999	1518 proteins	two reductions H,G -> H E,B -> E others -> C and H -> H E -> E others -> C	multiple alignment PSSM	window of 15	187 proteins divided into three sets	Low	76.5%,78.3%
Patttern Recognition Neural Network	PROF	Rost 2003	~2000 sequences	H,G,I -> H E -> E others -> C	multiple alignment PSSM	Outgrowth of PHD	201 proteins	Low	77.4%
Ensemble Model	EVA-4	Rost 2003	various	H,G,I -> H E,B -> E others -> C	multiple alignment PSSM	averages results from PSIPRED, PHDPSI,SSPRO, and PROF	201 proteins	Low	77.8%
Patttern Recognition Neural Network	SSPRO	Pollastri etal. 2002	1180 proteins	H,G,I -> H E,B -> E others -> C (SSPRO 2.0) none (SSPRO 8.0)	multiple alignment PSSM	Combines results from 11 Bidirectional Neural Networks context windows of 3 and 4	RS126; 223 EVA proteins; 40 CASP proteins	Low	77.7%; 78.13%; 80.65%
Patttern Recognition Neural Network	Petersen etal	Petersen etal 2000	1032 proteins	two reductions H,G -> H E,B -> E others -> C and H -> H E -> E others -> C	multiple alignment PSSM	Creates 800 predictions - one for each neural network configuration; weights the answers by reliability index; predicts i -1,i,i+1	RS126	Low	78.0%
Ensemble Model	PROTEUS	Montgomerie etal. 2006	1644 proteins from EVA	non DSSP program to assign secondary structure (VADAR)	multiple alignment PSSM	Combines results from PSIPRED, JNET and TRANSSEC	four tests: 100 proteins; 1644 proteins EVA 125 randomly chosen proteins; 10 random proteins	Low	77.6% wo homology search 81.3% with homology search

Table 8 - State of the Art Secondary Structure Prediction Efforts

2.2 Shannon's Information Theory

Information theory is a very important subject touching many areas including: statistical mechanics, communications, economics, decision analysis, and pattern recognition. As the application of information theory to protein structural uncertainty is central to the work presented here, this section will describe Shannon's landmark 1948 paper and then discuss some of the applications of information theory to protein science.

2.2.1 History

Interest in information, as an entity separate from the meaning of the message being conveyed grew out of work on telegraphs and telephones. The ideas of messages, channels, capacity, transmission rate, data compression codes and noise were all formally developed to better understand and implement long distance communication via telegraphs and telephones.

In 1948, Claude Shannon published a landmark paper entitled, "A Mathematical Theory of Communication". This paper, built on work by Nyquist and Hartley, developed several important ideas which form the basis of modern information theory. Chief among these is the idea of information entropy.

2.2.1.1 Information entropy

In his 1928 paper "Transmission of Information", R. Hartley had quantified information as follows:

$$H = \log_2 N$$

Where:

H = information
N = the number of possible events under consideration

This equation assumes that each of the N events is equally probable.

Shannon modified Hartley's formula to:

$$H = - \sum p(x) \log_2 p(x)$$

where H is information entropy.

This allows each of the N events, (x) to occur with a different probability.

In cases where the events are equally probable, for example a fair coin or die, Shannon's formula reduces to Hartley's information. In cases where events occur with different probability, say a secondary structure with probability of helix = .3; probability of coil = .5 and probability of sheet = .2, Shannon's would give a result where Hartley's would not.

Fair three faced die

$H = \log_2 (3) = 1.584963$ Hartley

$H = - [(1/3)*(\log_2(1/3)) + (1/3)*(\log_2(1/3)) + (1/3)*(\log_2(1/3))] = 1.584963$ Shannon

Secondary structure case

$H = - [0.3 * \log_2 (0.3) + 0.5 * \log_2 (0.5) + 0.2 * \log_2 (0.2)] = 1.48548$ Shannon

One can see that the uncertainty in the secondary structure case, with unequal probabilities, is less than that for the fair die with equal probabilities. This matches our intuition.

Shannon's equation also has a form similar to Boltzmann's H theorem for entropy.

$$H = \sum p(x) \ln p(x)$$

It is said that John von Neumann suggested to Shannon that he call his measure entropy because he recognized this correspondence. As a result, Shannon's H has been called information entropy.

2.2.1.2 Interpretations of H

H has many interpretations and uses. Shannon describes H as the entropy or uncertainty associated with a random event. It is also the minimum number of bits required to encode a message without information loss. Shannon used it to compute information loss in a noisy communication channel, the maximum transmission capacity of a channel, and the maximum degree of data compression for a given message and alphabet [Shannon, 1948]. The amount of information associated with a given observation is equal to the reduction in uncertainty associated with the event. This is sometimes called negentropy and is often used in statistics and pattern recognition studies.

$I = H_x$(before the observation Y) $- H_x$ (after the observation Y)
If x and y are independent, $I = 0$

2.2.2 Uses in Protein Science

Information theory has had many uses in protein science. Among these uses are: predicting secondary structure; measuring the effectiveness of predictors; measuring the effectiveness of different representations of proteins; predicting solvent accessibility and measuring evolutionary relationships. Examples of each of these will be discussed.

2.2.2.1 Predicting secondary structure

An early application of information theory to proteins is the Garnier, Osguthorpre and Robson or GOR algorithm. The basic idea is that conformation of a given a residue in a protein chain may be predicted using information from other residues in the same chain. The GOR method uses information theory to quantify the information gained by adding more residues to the mix.

> "The most general statement concerning the information we have for the conformation of the jth residue is thus
>
> $$I(S_j: R_1, R_2, \ldots R_{last}),$$
>
> which reads as the 'information which the first, second, and so on up to the last position carry about the conformation of the jth residue." [Garnier *et al.,* 1978 p 99].

Based on earlier research by the one of the authors, it was shown that most of the information about the conformation of a particular residue j was to be found in the eight residues on either side of j. Hence, the original GOR algorithm looks at a 17 amino acid window to predict the middle residue [Garnier *et al.,* 1978]. The GOR algorithm has continued to be improved in the thirty plus years since its introduction. The current version, GOR V, enjoys an accuracy of 73.5% [Kloczkowski, A. *et al*., 2002].

Ever since Rost and Sanders demonstrated the power of multiple sequence alignments (MSA) in the development of PHD, most predictors have taken advantage of them. Ding *et al.* attempted to predict secondary structure using a number of templates and a maximum entropy model while not using MSAs. A maximum entropy model selects that probability distribution which has the largest entropy consistent with known information.

Ding *et al.* start with CB513 and some sequences selected from EVA. They apply the Fishman – Argos eight to three reduction namely H-> H; E-> E; all others -> C. This is one of the least conservative reductions. resulting in as much as 3% apparent increase in Q_3. They divided the data sets into three classes: all-α, all β and α-β. The all-α class was more than 40% helix and less than 5% sheet. The all-β class was more than 40% sheet and less than 5% helix. The α-β class has more than 15% helix and more than 15% sheet.

They then used a set of class based propensity numbers patterned after Chou-Fasman and a set of feature templates based on amino acid properties and positions and a resizable window to predict secondary structure for each of the classes.

The results were compared to GORV using a jackknife test. The overall average of GORV on CB513 without MSAs was 67.5%. The numbers for Ding *et al.* are 77.4%, 73%, and 66.5% for all-α, all β and α-β respectively. When weighted by the number of proteins in each class the average is 69.7%. This represents an improvement of over 2% over GORV, the information theory secondary prediction pioneer [Ding *et al.*, 2009].

Crooks and Benner use information theory to measure the information in both the primary and secondary structures of both a large curated database they developed and CB513. They found that both the primary and secondary by residue entropy was quite high, 4.178 bits for the primary and 1.533 for the secondary. If they were completely random the numbers would be 4.322 and 1.585 respectively. They also found that while

the neighbor mutual information for the primary structure was quite low, 0.006 bits, the neighbor mutual information for the secondary structure was relatively high at 0.893 bits.

> "The inherent information content of secondary structure is 0.60 bits per residue, about four times greater than the 0.16 bits per residue of local mutual information between primary and secondary structure. These measurements put severe constraints on any single-sequence prediction algorithm that purports to extract secondary structure information from local sequence correlations."

Crooks and Benner then develop a hidden Markov model to predict the secondary structure. The accuracy ranges from 66.4% to 66.4% depending on the eight to three reduction used. This is comparable to the results from the original GOR. They then added information from multiple sequence alignments and the results improved to 72%, equivalent to the original PHD's results [Crooks and Benner, 2004].

2.2.2.2 Measuring the effectiveness of predictors

One of the uses of information theory in protein science is to measure the effectiveness of secondary structure predictors themselves. One of the problems with using Q_3 as a measure of merit for predictors is that it will sometimes overestimate the usefulness of the predictor badly. For example, assume we knew that the population of secondary structures for a particular database was 60% coil, 25% helix and 15% sheet. Any predictor which declared all of the structures to be coil would have a Q_3 of 60% and yet would have added nothing to our knowledge of the sequence. One method to avoid this overestimation is to focus on the reduction in uncertainty (additional information) provided by each algorithm and select those which result in the most information.

Swanson *et al.* [2008] have developed a method to use information theory to measure the effectiveness of secondary structure prediction over time. They calculate the observed and predicted entropies for secondary structure in the normal way. They then calculate the entropy of the joint distribution (%cc, %ce, %ch, %ec, %ee, %eh, %hc, %he, %hh) and subtract it from the sum of the observed and predicted entropies. This difference gives the amount of information provided by the prediction algorithm. These numbers are normalized for the entropy in the observed and reported for the most successful predictors in the first five CASP contests. CASP is Critical Assessment of Techniques for Protein Structure Prediction. It is a bi-annual contest for objectively testing computer models in protein structures. The first five contests starting in 1994 has a secondary structure prediction component. What Swanson *et al*. showed was that the best predictors were getting better not only on a Q_3 basis, growing from 72 to 81% but also from an information theory perspective. The % of available information provided nearly doubled, from 27 to 51%.

2.2.2.3 Measuring the effectiveness of representations

Katzman *et al.* use information theory to compare different alphabets for use in prediction algorithms. The basic idea is to use different alphabets to depict different characteristics of a residue which may allow for better predictions. Building on work by Karchin and coworkers, Katzman *et al.* evaluated several alphabets used to describe primary and secondary structure. Some the properties they looked at were participation in hydrogen bonding, torsion angles, and accessibility to solvents.

Each was run through two neural network architectures and the information gained was computed. This allowed each of these very diverse alphabets to be compared. The greatest information gain was attained by using the str2 alphabet. STR is a modification of DSSP where the letter E is divided into six letters depending on how it interacts with its strand neighbors (parallel, anti parallel etc,). The str2 alphabet gained over 1 bit in both architectures but had a middling Q_3 of (0.54-0.56)

Katzman *et al.* state that the results of work allow them to use backbone based alphabets (DSSP,STRIDE etc) for secondary structure prediction and other alphabets for cost functions and tuning three dimensional models[Katzman *et al.*, 2008].

Another study which uses information to evaluate the effectiveness of representations is Zhang *et al.* [2008]. K. C. Chou, one of the authors, has developed the pseudo-amino acid (PseAA) [Chou, 2001] representation for proteins and another called functional domain (FunD) composition. The PseAA is a list of the 20 amino acids relative frequencies followed by any number of weighted sequence based correlations (i+1,i+2,etc.). The functional domain composition is based on a database of functional domains called InterPro [Chou and Cai, 2004]. Each of the 7785 domains is an element in the vector describing a protein. Chou then uses this representations to predict the class (all-α,all-β, α/β, etc.) membership of each protein. The success reported is nearly 100%.

Zhang *et al.* recognize that PseAA representations can result in overfitting and that the large information requirements of the FunD composition may be unavailable. To address

this they developed a PseAA which incorporated hydrophobicity and approximate entropy. They used a fuzzy K nearest neighbor classifier to predict which class, all-α,all-β, α+β, α/β etc., each protein belongs to. They report an overall success rate of 97% [Zhang *et al.*, 2008].

2.2.2.4 Predicting solvent accessibility

Naderi-Manesh *et al.* developed an algorithm using information theory to predict when a side chain would be accessible to a solvent/water. It is modeled after the GOR program. Like the GOR program, Naderi-Manesh *et al.* use data from the eight residues on either side of an amino acid to predict accessibility. They define different levels of accessibility from 5% to 81%. For the three state case, buried, intermediate , and exposed, they report an accuracy of approximately 60% depending on which threshold is used. This on the same order as the figures for the original GOR predictions for the three state case [Naderi-Manesh *et al.*, 2001].

2.2.2.5 Evolution

Martin *et al.* use mutual information to identify residues which are co-evolving. By this they mean residues which are not conserved across multiple alignments of homologous sequences but where functionality is conserved by forcing mutations in two or more residues at the same time.

Mutual information is defined as:

$MI(X,Y) = H(X) + H(Y) - H(X,Y)$

where

$H(X,Y) = - \sum\sum p(x,y) \log_b p(x,y)$

and b is any arbitrary base.

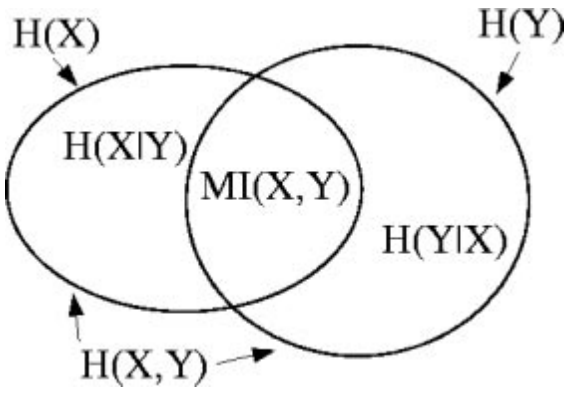


Figure 15 – Venn Diagram
Martin et al., 2005

They develop and test their technique on a simulated evolution and then apply it to real sequences. After normalizing and ranking the amino acid pairs by mutual information they discovered that isolated pairs were often in close physical contact even if they were widely separated in the sequence order and that high ranking networks (more than one partner) were often near active sites [Martin *et al*., 2005].

2.2.2.6 Summary

Shannon's 1948 paper established the field of information science. It has proven foundational to many different areas including communications, pattern recognition, economics, and statistics. It has also been used to great effect in the study of proteins.

2.3 Chameleon Sequences

2.3.1 Definition

Structurally ambivalent sequences (SAS) are sequences of amino acids that have been identified with different secondary structures. Chameleons are the extreme case of structurally ambivalent sequences. When Minor and Kim first coined the term it referred to a specific engineered 11 amino acid sequence which formed a helix under some conditions and a sheet under others. The current definition of a chameleon sequence is a sequence of amino acids which have a helix secondary structure in one protein and an extended secondary structure in another. Several researchers have investigated chameleons and SASs.

2.3.2 Kabsch and Sander

Kabsch and Sander reviewed 62 proteins and found that the longest sequences which were part of a helix in one protein and a sheet in another protein were five amino acids long (a pentapeptide). They did find two six amino acid sequences outside their database (CRNKAS and GYITDG). Twenty-five such sequences were found. They found seven sequences which exemplified “same sequence-different structure” and compared them to six pentapeptides which illustrate “same sequence-same structure”. They showed that one cannot predict secondary structure on the basis of sequence similarity alone, particularly for such short sequences [Kabsch and Sander, 1984].

2.3.3 Cohen *et al.*

Cohen *et al*. reviewed the July 15, 1990 version of the PDB. Starting with 366 sequences from 315 proteins, they identified 59 pairs of hexapeptides. Of these, eight pairs had a

helix structure in one protein and a sheet structure in another. In none of the eight did the members of the pair share the same SCOP folding class (α, β, α/β, α+β). Cohen *et al*. ran each chameleon through a program (CONFORM) which implemented the Chou-Fasman rules. Their data shows that the results are consistent with the 55-66% accuracy associated with Chou-Fasman predictions, but this is on a very small data set (16) with $3/8^{ths}$ of the data (6) giving no meaningful prediction [Cohen *et al*.,1993].

2.3.4 Kim and Minor

Kim and Minor identified an engineered sequence, eleven amino acids long, which they inserted into two different places within a single protein. In one spot the sequence folded into a helix, in another into a sheet. In this way they showed that secondary structure was determined not only by the local sequence but also by its environment. They also coined the term 'chameleon' to describe a peptide which has the same primary sequence but a different secondary structure depending on circumstances [Minor and Kim, 1996].

2.3.5 Sudarsanam

Sudarsanam reviewed the April 1996 PDB and found that four pairs of octamers with different secondary structure and eight pairs of heptamers with different secondary structure. None of these were helix–extended chameleons [Sudarsanam, 1998].

2.3.6 Mezei

Mezei defined chameleon sequences as only those which were completely sheet or completely helix. This definition has generally been adopted. He then reviewed the April 1997 PDB and identified three that were seven residues long, thirty-eight which

were six residues long and nine hundred-forty which were five residues long [Mezei, 1998].

2.3.7 Zhou *et al.*

Zhou *et al.* reviewed the June 1999 PDB. They used STRIDE to assign secondary structure. Two databases one with less than 25% sequence identity and one with less than 95% identity were created. Zhou *et al.* then looked for sequences which had either partial or complete helix to sheet transition. In the first dataset they found seventy-three 7-mer pairs of which 16 had partial transitions and none had complete transitions. In the second database they found one thousand-nine-hundred and thirty two (1932) 7- mer pairs with 86 partials and 2 complete transitions. They then took one hundred-sixty-seven (167) tetramers which were strongly ambivalent and compared the predictions produced by PHD. For these, serious errors were made in 13.2% of the cases. This compared for 5.3% for other tetramers and around 8% average confusion between helix and strand for all residues.

Zhou *et al.* calculated the normalized frequencies for all dipeptides found in n-mers with complete helix to strand transitions. Eight of the ten most frequent had a strong helix former and strong strand former (Chou-Fasman) coupled together. They interpreted this as evidence that these dipeptides had a large degree of inherent local flexibility, allowing the global environment to determine the final secondary structure. This was supported by the fact that they found that the 4, 5, and 6-mer pairs with complete helix to strand transition were overwhelmingly from different SCOP classes, 82.2%, 85%, and 75% respectively. Numbers are similar for partial transitions [Zhou *et al.*, 2000].

2.3.8 Jacoboni *et al.*

Jacoboni *et al*. [2000] worked to answer the question, " …to what extent predictors are able to distinguish the different structures of chameleon sequences." To this end, a large database of chameleons was built of 2576 pairs of chameleons 5-8 amino acids in length from 755 proteins of < 25% similarity.

To explore structural diversity and its effect on secondary structure prediction, Jacoboni *et al*. defined a chameleon a little differently than others. For Jacoboni a chameleon was any sequence where the secondary structure differed at every position. Hence, CCCHHH and EEECCC would qualify here where it would not for others (Section 2.2.6).

Using their database, Jacoboni *et al*. tested several secondary structure predictors including: GORIV, PHD, PSI-Pred, JPRED, PRED2ARY, DSC, NNSSP and PREDATOR. They also developed their own neural network predictor to test the sensitivity of the predictions to different multiple alignment algorithms.

Jacoboni *et al*. showed that the first and second generation techniques using a single sequence did significantly worse when predicting chameleons. Third generation techniques such as PHD, PSI-Pred and JPRED which use multiple alignments as an input did nearly as well on chameleons as they did on typical sequences. Accuracy was 72-78% for typical sequences and 72-75% for chameleon sequences. Single sequence predictors fared 7.6% worse on chameleons when compared to the general accuracy.

2.3.9 Kuzenetsov and Rackovsky

Kuzenetsov and Rackovsky developed a measure they call a generalized local propensity (GLP) to measure conformational uncertainty of a tripeptide (an amino acid and its two neighbors). They did this by computing the relative Shannon entropy of the distribution of dihedral angles for the central residue of a tripeptide. "A positive value of $glp({}_iX_j)$ indicates that the amino acid type X in the tripeptide ${}_iX_j$ has conformational variability that is lower than average. A negative value…indicates… higher conformational variability than average."

They then compared these measurements for five different classes of structurally ambiguous fragments. They are: Helix-Extended (0.43), Helix-Irregular (0.30), Extended-Irregular (0.153) Irregular-Irregular (0.101), and Mixed (not available). They also studied the amino acids flanking the tripeptides.

> "The flanks of chameleon k-mers in helical and sheet conformations show the greatest difference in local propensity. …Chameleon hexamers occur in the helical conformation 59% of the time, whereas only 15% of chameleon hexamers in the β-sheet conformation are located in the middle of a β-sheet. We conclude that strong local helical propensity in the flanking residues forces chameleon k-mers to adopt a helical conformation. In the absence of flanks with high local coding propensity, chameleon k-mers adopt the more energetically favorable extended conformation." [Kuznetsov and Rackovsky 2003]

2.3.10 Tankano *et al.*

Tankano *et al.* investigated a particular chameleon sequence, namely TQDMINKST. It is one of the very rare sequences that take on both conformations in the same protein (MATα2/MCM1/DNA). When this sequence was attached to a helix in another protein it

took on a helix structure. When attached to an extended structure in a third protein, it took on an extended conformation. When it was attached to a protein outside a helix or sheet region it took on no structure at all. They coined the term conformational contagion to describe this phenomenon [Tankano *et al.*, 2007].

2.3.11 Guo *et al.*

Guo *et al.* looked at the question of whether chameleons are more difficult to predict than other proteins. Some researchers found that chameleons may pose a problem for prediction [Zhou *et al*., 2000]. Guo *et al.* defined two types of chameleons. The first is HS or helix-strand. The second is HE, or helix-sheet. They found two eight-residue and 56 seven-residue HS sequences. They found 7 seven- and 39 six-residue HE sequences.

Guo *et al.* computed the relative frequency of each of the amino acids in the chameleons and their flanks. They found that V, L, I and A have the highest relative frequency in chameleon-HS sequences. They observe that A and L have a strong propensity for helix while V and I have a strong β sheet propensity. C, H, M, P and W occur much less frequently.

Guo *et al.* also tested whether chameleons were more difficult to predict than non chameleons. They used both Chou-Fasman and PsiPred to predict secondary structure for the chameleon sequences under study. These represent a first and third generation predictor respectively. They found that both predictors predicted helices in chameleon-HS sequences(≥6) more accurately than for proteins in general (95% vs. 85.4% and 61.1% vs. 56.1%) and that PsiPred predicted strands more accurately (81.4% vs 78.6%)

while Chou-Fasman was less accurate on chameleon HS strands (39.2% vs. 46.7%) [Guo *et al.*, 2007].

2.3.12 Chameleon Sequence Summary

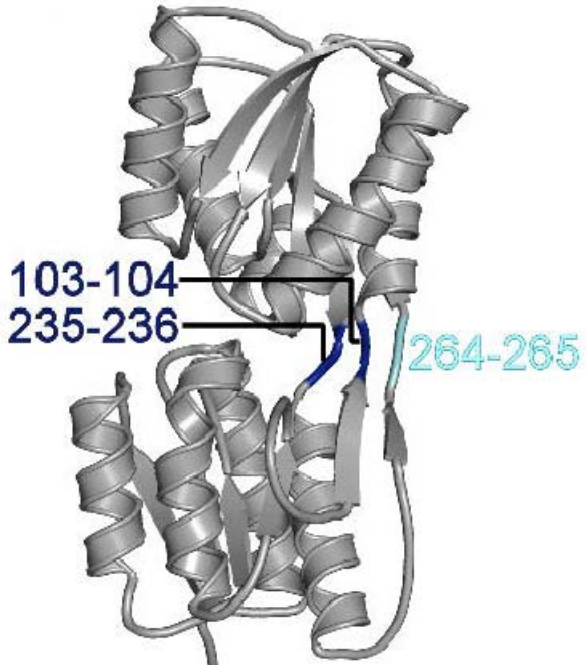
The chameleons are short (≤ 8 amino acids) and rare. The literature is mixed on the effects of chameleons on secondary prediction. Earlier investigators state that chameleons are particularly difficult to predict while later researchers claim that 3rd generation methods are not affected by chameleons.

2.4 Protein Hinges

2.4.1 Importance of Hinges

> "It is common to classify protein movements into hinge and shear. Hinge movements involve rotation of protein parts (mostly domains) about a region called a hinge (in most cases a loop or a linker). This region usually involves several residues that undergo significant conformational changes, but most of the rotating protein parts remain unchanged. Shear movements involve a sliding movement of protein parts relative to each other. This movement usually restricted, with small conformational changes across the movement interface plane."[Emekli, 2007]

Detecting hinges is a critical part of understanding the relationships between a protein's structure and function. A number of researchers have been developing methods to identify hinge regions and their associated domains. Below are discussions of a number of representative papers in this area.



Ribose Binding Protein (1URP)
[Keating, et al. 2009]

Figure 16 - Protein Hinges

2.4.2 HingeFind

Wriggers and Schulten [1997] developed a method to identify protein hinges and domains called HingeFind. It is based on comparing two conformations of the same sequence and identifying those portions which can rotate. It starts with the two conformations and lines up the corresponding carbon atoms. Candidate domains are iteratively grown until an error threshold is met. Then hinges are identified and rotations are tested. HingeFind was then tested on four proteins resulting in general agreement with the literature.

2.4.3 FlexProt

Shatsky *et al.* [2004] start with two sequences and then

> "…decompose the two molecules into a minimal number of disjoint fragments of maximal size, such that the matched fragments will be almost congruent. We define two fragments to be almost congruent if their sequence lengths (measured by the number of Cα- atoms) are the same and there exists a 3-D rotation and translation which superimposes the corresponding atoms with a small RMSD."

This is shown in Figure 17. The overlap region is a candidate hinge.

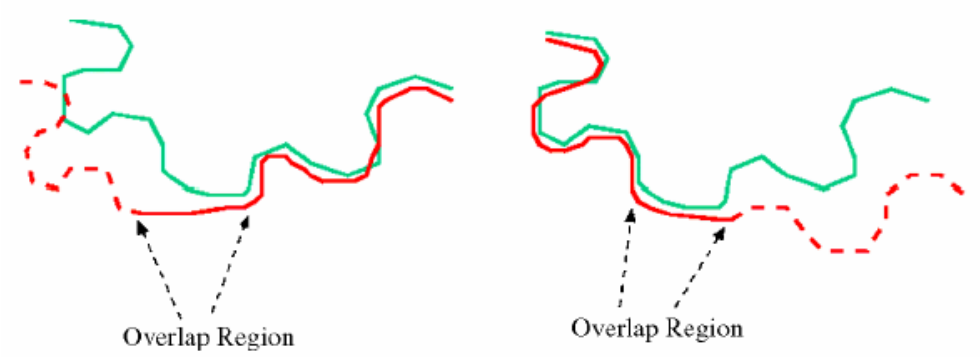


Figure 17 - Overlap Regions

Shatsky *et al.* discuss the computational complexity time associated with their algorithm. They show that the theoretical limit is $O(n^6)$ but that through a number of algorithm modifications and transformations associated with their implementation the complexity is reduced to $O(n^4)$. The algorithm is then tested on four pairs of sequences. The superimpositions and RMSDs are given in the paper but there is no quantitative comparison to the findings of other researchers on the same sequences.

2.4.4. Hinge Atlas

Flores *et al.* [2007] reviewed a number of hinge motions from the Database of Macromolecular Motions. [http://www.molmovdb.org] Key findings include that glycine and serine are more likely to occur in hinges. Phenylalanine, alanine,valine and leucine are less likely to appear in hinges. Hinges are often found in random coil. If one uses the eight to three reduction of H, G -> H; E, B -> E; all others -> C; Flores *et al*. report that hinges occur 67% of the time in coil, 21% of the time in sheet and 12% helix.

Flores *et al.* also found that hinges are often within four residues of an active site. They developed an index they call HingeSeq, incorporating the amino acid, secondary structure, and distance from an active site. While the index was highly statistically significant, it was not sensitive and therefore, a poor predictor of hinges.

2.4.5 HingeProt

HingeProt is an automated predictor of hinges developed by Emekli *et al.* [2008]. It is based on an Elastic Network Model. In an elastic network model the protein is envisioned as a set of balls (alpha carbons) and springs. The normal modes of the system are then computed and the flexibility of different regions is calculated. In this way rigid and hinge areas are identified. Emekli *et al.* compared their method with another predictor, FlexProt (Section 2.4.3). FlexProt requires two conformations, HingeProt only one. HingeProt also had better coverage and a better alignment (lower root mean square distance (RMSD)) than FlexProt for the proteins tested (89% vs 70% coverage and 1.4 Å vs 1.5 Å RSMD). The HingeProt server is at http://www.prc.boun.edu.tr/appserv/prc/hingeprot/, (Dec 22, 2010).

2.4.6 StoneHinge

StoneHinge is an automated detection method which uses the consensus of two network based models, StoneHingeP and StoneHingeD, to predict the existence of a hinge. StoneHinge, like HingeProt, only uses one conformation of a protein. Other predictors use two conformations (open and closed) to identify hinges by finding those portions which move. StoneHingeP is built on a model called ProFlex which counts constraints in three dimensions evaluates whether a particular region is rotatable or not. StoneHingeP takes the results of ProFlex and identifies areas which may be rigid (domains) or flexible (hinges).

StoneHinge D uses a Gaussian Network Model similar to HingeProt to find hinges and domains. The results of the two models are compared and where they agree within five residues, the results are combined and reported.

Keating *et al.* [2009] tested the models against twenty protein structures. Nine proteins had eighteen hinges. Of these, thirteen were identified in the same place reported in the literature. Most of the correct predictions were in the open conformation.

2.4.7 Fast Hinge Detection Algorithms

Most hinge detection methods rely on comparing two conformations of the same protein sequence and identifying portions which may have moved or stayed the same. The usual metric used is the root mean square deviation or (RMSD). Shibuya [2010] has developed a measure which extends RSMD to include hinges. He assumes that the hinge is a single atom and then rotates each atom in turn to find a minimum RMSD. Hence, the value can

be computed in linear time. He extends this idea to *k* hinges within a single sequence. He then uses his computed $RMSDh^{(k)}$ to predict the number and position of the hinges. He requires each rigid fragment to be at least 15 residues long. He then finds all of the fragments where the RMSD is below a given threshold and declares the position of the hinges which connect these fragments.

He tested these algorithms on twelve proteins using 1.5 Å as the threshold. He correctly predicts the number of hinges in 9 out 12 cases and the positions of the hinges in six of twelve proteins. This compares favorably with FlexProt which achieved four and one on the number and position of hinges respectively.

2.4.8 Protein hinge summary

Hinges are critical to the functioning of proteins. There are several methods to identify proteins and the domains they join. Most of the methods require two conformations (open and closed) to compare. Flores *et al.* have developed a Hinge Atlas which gathers several hinge sequences into one place along with associated data for use by the larger community.

3.0 METHOD DEVELOPMENT AND APPLICATION

3.1 Using Shannon's H as a Uncertainty Measure

3.1.1 Motivation

After reviewing the literature on chameleons, hinges and secondary structure it became apparent that a method to directly quantify the uncertainty of secondary structure in response to a given primary structure is needed. Chameleon sequences represent the most extreme example of this uncertainty. This work develops and uses such a method to answer questions about secondary structure chameleon sequences and protein hinges.

3.1.2 Design goals

The design goals for the method include: 1) quantifiable; 2) easy to use; 3) easy to understand; 4) able to identify things which are the same; 5) distinguish between things which are different; 6) robust, in the sense that minor differences do not give wildly different results; 7) consistent; 8) and scalable, with results comparable across many sizes of items.

3.1.3 Candidate Method

A number of potential methods for quantifying uncertainty of secondary structure were explored. One such method, based on computing the difference between the number of helices and sheets for each position-amino acid looked promising. Unfortunately, this technique was unable to distinguish sequence windows found to be equally frequent in helices and beta strands, thus failing design goal five above. For example, a case with

helix .33, coil .34, sheet .33 has the same measure (0.0) as another case with helix 0.5, coil 0.0 and sheet 0.5 (0.0). Yet the first case clearly has more uncertainty than the latter.

3.1.4 Method to Quantify Uncertainty

Shannon's measure of information entropy overcomes the difficulties encountered in the helix/sheet counting technique mentioned above. In fact, it appears to meet all of the design goals outlined in Section 3.1.2. As a result, the method outlined below is based on Shannon's H.

The method is as follows:

1. Select a reference set
2. Select a window size (3, 5, 7 etc.)
3. Compute the uncertainty value H for the secondary structure associated with the central amino acid within all windows of the selected size in the reference set. Store it in a lookup table.
4. For each sequence to be measured, look up the uncertainty measure in the table for each amino acid in the sequence.

3.1.5 Reference Set

The reference set selected is critical to this analysis. The database selected for this study is CB513. Developed by Cuff and Barton to train their secondary structure prediction model JPRED, CB513 is a collection of 513 sequences. Constructed to carefully control homology, CB513 is readily available at http://www.compbio.dundee.ac.uk/www-jpred/about.html, (Dec 22, 2010). As a result, it has been used by numerous researchers

studying secondary structure. It has 84,119 amino acids which have the following distribution:

A	7267	I	4642	S	5222
B	31	K	4976	T	5015
C	1381	L	7134	V	5795
D	4973	M	1710	W	1236
E	5050	N	3976	X	19
F	3268	P	3903	Y	3065
G	6657	Q	3108	Z	14
H	1865	R	3812	Total	84119

Table 9 - CB 513 Distribution of Amino Acids

The non standard amino acids B, X, and Z account for only 64 residues out of the 84,119. The secondary structure was reduced using the following eight to three mapping:H,G -> H; E,B-> E; all others to C. Following this, the secondary structure counts were

C	35993	0.43
E	19059	0.23
H	29067	0.34
Total	84119	1.00

Table 10 - CB513 Secondary Structure

3.1.6 Class, Architecture, Topology, and Homology (CATH)

Another way to characterize protein sequences is by their tertiary structure. One method to do this is via the CATH hierarchy. Developed by Orengo *et al*., the Class, Architecture, Topology, and Homology (CATH) database is a curated hierarchical collection of protein sequences from the PDB. At the top level of class it has four classes: mainly alpha; mainly beta; mixed alpha-beta; and few secondary structures.

Each of these is divided into architectures. Architectures describe the overall shape of the domain structure. It contains some well known motifs, *e.g.* beta barrels *etc*. The architectures are then divided into topologies according to their folds. The topologies are further separated into homologous super families. The super families are then ordered by the level of sequence identity and overlap. Each domain is given a CATH number depicting where it sits in the hierarchy. The number of domains in each class and architecture is listed in Appendix F [Orengo *et al.* 1998].

In order to characterize the data bases used in this study, each sequence was reviewed and its CATH number was recorded. The distribution of domains found within CB513 is given in Table 10.

Table 11- Distribution of CB513 Protein Sequences by CATH Architecture

1.0	Mainly Alpha	
1.10	Orthogonal Bundle	64
1.20	Up-down Bundle	24
1.25	Alpha Horseshoe	1
1.40	Alpha solenoid	0
1.50	Alpha/alpha barrel	3
2.0	Mainly Beta	
2.10	Ribbon	8
2.20	Single Sheet	3
2.30	Roll	5
2.40	Beta barrel	28
2.50	Clam	1
2.60	Sandwich	61
2.70	Distorted Sandwich	5
2.80	Trefoil	4
2.90	Orthogonal Prism	0
2.100	Aligned Prism	2
2.102	3-layer Sandwich	1
2.105	3 Propellor	0
2.110	4 Propellor	1
2.115	5 Propellor	0
2.120	6 Propellor	2
2.130	7 Propellor	2
2.140	8 Propellor	0
2.150	2 Solenoid	0
2.160	3 Solenoid	1
2.170	Beta Complex	0
3.0	Mixed alpha-beta	
3.10	Roll	24
3.15	Super Roll	0
3.20	Alpha-Beta Barrel	25
3.30	2-Layer Sandwich	76
3.40	3-Layer(aba) Sand.	99
3.45	3-Layer(aab) Sandwich	0
3.50	3-Layer(bba) Sandwich	6
3.55	3-Layer(bab) Sandwich	0
3.60	4-Layer Sandwich	4
3.65	Alpha-beta prism	0
3.70	Box	0
3.75	5-stranded Propellor	0
3.80	Alpha-Beta Horseshoe	1
3.90	Alpha-Beta Complex	20
3.100	RibosomalProtein L15; Chain K; domain2	0
4.0	Few Secondary Structures	
4.1	Irregular	11
	Not assigned	31
	Total	513

3.1.7 Analysis Overview

Shannon's H will be computed for the secondary structure of each central amino acid within a window of three using CB513 as the reference set. The regions of interest, chameleon or hinge, will be identified along with their respective flanking regions. The average information-entropy or uncertainty for each region is then computed. These averages are compared among the five regions (outer N terminus flank, N terminus flank, chameleon or hinge, C terminus flank, outer C terminus flank) using a T-test.

This is then repeated for each case substituting Chou-Fasman numbers for the Shannon H values. Since there are three sets of Chou-Fasman numbers(Pa, Pb, Pturn), there are three sets of analyses. The results are then compared.

3.2 Application – Chameleons

3.2.1 Hypotheses

The definition of chameleon sequence adopted for this effort is two amino acid sequences which have the same primary structure and a helix secondary structure in one case and an extended secondary structure in another case. It is believed that chameleon sequences take their secondary structure from their neighboring amino acids. If this is true, one would expect that the uncertainty of the flanking regions would be less than normal since they would be more rigidly helix or sheet. The uncertainty of the chameleon regions, on the other hand, would be greater than normal to reflect its presumed greater flexibility. From this come three sets of hypotheses:

1-H_0: $U_{cham} = U_{flank}$	2-H_0: $U_{flank} = U_{other}$	3-H_0: $U_{cham} = U_{other}$
1-H_1: $U_{cham} \neq U_{flank}$	2-H_1: $U_{flank} \neq U_{other}$	3-H_1: $U_{cham} \neq U_{other}$

The first of these null hypotheses say that there is no difference between the uncertainty measure computed for the chameleon regions and their flanking regions. The second, that there is no difference between the U computed for the flanking region and other regions in the protein. The third, that there is no difference between the U computed for chameleon regions and others in the protein. Each of these will be investigated in turn.

3.2.2 Data Sets

The choice of data sets was a key to this effort. The first choice was the use of CB513 as the reference set. The second dataset used in this investigation was ss.txt. It contains the sequence and secondary structure of most of the proteins in the Protein Data Bank in a slightly modified fasta format. An example is given below. This dataset is available at http://www.rcsb.org/pdb/files/ss.txt. I used the ss.txt file available March 20, 2009.

```
>1TNF:A:sequence
RTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLL
THTISRIAVSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLLFAES
GQVYFGIIAL
>1TNF:A:secstr
CCCCCCCEEEEECCCCCCCCCCCCCCCCCEECCCCEECCCCEECCCCEEEEEEEEEEEEECCCCCCCCCE
EEEEEEECCCCCCEEEEECCEECCCCCCCCCCCCCCEEEEEEEEEEEEEECCCCEEEEEECCHHHECCCCC
CCCEEEEEEC
```

Figure 18 - ss.txt Format

CB513 was converted to this format and the eight to three reduction H,G,I => H; E,B => E; all others to C was used to reduce the secondary structure.

3.2.3 Chameleon Database Development

The next step was to develop a database of chameleons. This was accomplished in three major steps.

1. Find chameleons
2. Validate data
3. Control homology

3.2.3.1 Find chameleons

A ruby script was written to identify all chameleons of a given length within the ss.txt database. This identified several sequences which were dramatically longer than any chameleons previously reported. The longest reported naturally occurring chameleons are eight amino acids long. Each of these very long sequences were then compared to their entries in the larger Protein Data Bank and found to be problematic in one or more ways.

3.2.3.2 Validate data

Following this all of the candidate chameleon sequences were checked against the PDB using the following criteria.

1. Both proteins which contain the sequence are in the PDB (not obsolete or replaced).
2. The proteins are real (not a theoretical model).
3. The proteins are resolved using X-ray crystallography.
4. The proteins have resolution of 4 angstroms or better.
5. The chameleon appears in both proteins primary sequences.

6. The chameleon secondary structure agrees with the author approved secondary structure.

Using these criteria, the problematic proteins were removed from the database. The find chameleon script was then rerun against the reduced data set to produce a list of validated chameleons.

3.2.3.3 Control Homology

The last major step in building the chameleon database was to control homology. This was done by running the database against itself using the Basic Local Alignment Search Tool (Section C.2.2 BLAST). Any proteins with more than 20% similarity were then eliminated. This included both those which were identical and those which conserved physiochemical properties.

3.2.3.4 New Chameleons

As a result of these efforts several new chameleon sequences were identified. Nine helix/extended (H/E) chameleons of length eight and eighty-five (85) chameleons of length seven were found. Prior to this effort the greatest number of naturally occurring H/E chameleons reported in the literature were two of length eight and sixty-three (63) of length seven. (Guo *etal*. HS and HE results are combined.) The CATH number distribution is also included.

Table 12 - Chameleons of Length Eight

SEQUENCE	PROTEIN	POSITION	SECONDARY	PROTEIN	POSITION	SECONDARY
AVRLAALN	2BKU:B	195	HHHHHHHH	2E3V:A	80	EEEEEEEE
ELVKLVTH	1BXI:A	31	HHHHHHHH	3E0G:A	131	EEEEEEEE
FALDLLME	1R2F:A	220	HHHHHHHH	1YQ3:A	176	EEEEEEEE
KSLLDYEV	2RCA:A	222	HHHHHHHH	2R5R:A	123	EEEEEEEE
SAVVLSAV	1GW5:B	252	HHHHHHHH	2J7N:A	109	EEEEEEEE
SVTAFLND	1PT6:A	24	HHHHHHHH	2A6X:A	123	EEEEEEEE
VEGRAILR	3F13:A	117	HHHHHHHH	3CMB:A	267	EEEEEEEE
VITAGIGI	1EYS:M	275	HHHHHHHH	2Z1E:A	143	EEEEEEEE
VLYVKLHN	1I3P:A	275	HHHHHHHH	1JIG:A	18	EEEEEEEE

Table 13 - Chameleons of Length Seven
(includes 7s which are part of 8s)

SEQUENCE	PROTEIN	POSITION	SECONDARY	PROTEIN	POSITION	SECONDARY
AAAVFHN	1KHV:A	200	HHHHHHH	1SCI:A	97	EEEEEEE
AAIVGAA	1UM9:B	60	HHHHHHH	1XN1:C	70	EEEEEEE
AIAAVTV	1J22:A	99	HHHHHHH	3C8L:A	110	EEEEEEE
AIQVLPK	2R9R:A	301	HHHHHHH	2NQO:D	165	EEEEEEE
ALATRLV	1MVM:A	64	HHHHHHH	3EF6:A	373	EEEEEEE
ALRAVTT	1GT8:A	773	HHHHHHH	2YWC:D	448	EEEEEEE
ATVAALA	1H0H:A	160	HHHHHHH	1QHV:A	46	EEEEEEE
AVIESVV	2VLB:A	51	HHHHHHH	2UV8:G	1375	EEEEEEE
AVRLAAL	2BKU:B	195	HHHHHHH	2E3V:A	80	EEEEEEE
AVVLSAV	1GW5:B	253	HHHHHHH	2J7N:A	110	EEEEEEE
DKFLVLA	1U0L:A	104	HHHHHHH	2PNQ:A	338	EEEEEEE
DLTIKLV	1IRU:D	184	HHHHHHH	1GKU:B	754	EEEEEEE
EDKLVVH	1SKY:E	394	HHHHHHH	1ROW:A	70	EEEEEEE
EESRTEV	1YWM:A	143	HHHHHHH	1KSI:A	362	EEEEEEE
EFIAAVN	2GGZ:A	77	HHHHHHH	1JVN:A	179	EEEEEEE
EGRAILR	3F13:A	118	HHHHHHH	3CMB:A	268	EEEEEEE
EITFLKN	1FBM:A	30	HHHHHHH	2P8G:A	129	EEEEEEE
EKALELV	1R89:A	7	HHHHHHH	1C04:B	82	EEEEEEE
ELRLMVA	1P16:A	23	HHHHHHH	2AE0:X	288	EEEEEEE
ELSARYA	1V6S:A	125	HHHHHHH	3G7G:A	71	EEEEEEE
ELTLSIT	2PLW:A	130	HHHHHHH	1YLN:A	40	EEEEEEE
ELVKLVT	1BXI:A	31	HHHHHHH	3E0G:A	131	EEEEEEE
EMAVAAA	2DG6:A	181	HHHHHHH	2POR:A	148	EEEEEEE
ESVLVGA	1ZCH:A	104	HHHHHHH	2P9W:A	245	EEEEEEE
EVEEGLA	1H0C:A	137	HHHHHHH	2J0W:A	380	EEEEEEE
EVTKVMA	1V4N:A	227	HHHHHHH	2PN2:A	93	EEEEEEE
FEAAIAA	1PSQ:A	152	HHHHHHH	3E39:A	144	EEEEEEE
FGAVGAL	2I7N:A	348	HHHHHHH	2A8I:A	126	EEEEEEE
FLEGFVR	1SFR:A	228	HHHHHHH	1XF1:A	579	EEEEEEE
FSAMTSA	1XJT:A	105	HHHHHHH	1YDG:A	117	EEEEEEE
FSVTGNV	1M6D:A	27	HHHHHHH	1M06:G	17	EEEEEEE
FYSVVEL	1TFF:A	101	HHHHHHH	1VDH:A	107	EEEEEEE

Chameleons of Length Seven
Continued

SEQUENCE	PROTEIN	POSITION	SECONDARY	PROTEIN	POSITION	SECONDARY
GAILSLS	1KG2:A	122	HHHHHHH	1K0G:B	249	EEEEEEE
GEVEALV	1B70:B	707	HHHHHHH	3BHD:A	174	EEEEEEE
GFLVTIK	1C4T:A	213	HHHHHHH	2HQL:A	27	EEEEEEE
GGILATA	1SZQ:A	117	HHHHHHH	1WXW:A	317	EEEEEEE
GRVGVAA	2DVL:A	234	HHHHHHH	2QTK:A	92	EEEEEEE
GSGILAL	1IO0:A	106	HHHHHHH	3CSL:A	259	EEEEEEE
GTLVGLA	1KPK:A	42	HHHHHHH	1UYN:X	205	EEEEEEE
GVTNKVN	2FK0:B	46	HHHHHHH	2QV3:A	135	EEEEEEE
HADIQVR	2ZIE:A	115	HHHHHHH	1HWH:B	149	EEEEEEE
IAQLTVN	2F1M:A	77	HHHHHHH	3DWO:X	43	EEEEEEE
IDAASIA	2NN6:A	162	HHHHHHH	1WUB:A	49	EEEEEEE
IFVTLLI	1M56:B	44	HHHHHHH	2OZP:A	171	EEEEEEE
IKMFIKN	1NSJ:A	194	HHHHHHH	1PGS:A	42	EEEEEEE
IRQIFAL	2ON5:A	17	HHHHHHH	1JJU:A	239	EEEEEEE
KICSIAL	2AJ4:A	469	HHHHHHH	2V4U:A	23	EEEEEEE
KKSAKTT	2FEZ:A	262	HHHHHHH	1GMN:A	15	EEEEEEE
KLIAIKM	2VIX:A	182	HHHHHHH	1XIQ:A	38	EEEEEEE
KSLLDYE	2RCA:A	222	HHHHHHH	2R5R:A	123	EEEEEEE
KVYNALR	2VSQ:A	163	HHHHHHH	2QPV:A	103	EEEEEEE
LEFYYDK	1EYU:A	119	HHHHHHH	3CSL:A	458	EEEEEEE
LESVEFW	2PD0:A	208	HHHHHHH	1T9M:A	180	EEEEEEE
LGIALSH	1VKW:A	182	HHHHHHH	1RM6:A	436	EEEEEEE
LPVLVRQ	2ZOP:A	34	HHHHHHH	2YWD:A	156	EEEEEEE
LTELFVK	1OU5:A	61	HHHHHHH	1GT1:A	114	EEEEEEE
LTVRAAR	1G8F:A	206	HHHHHHH	1WWL:A	72	EEEEEEE
LVKLVTH	1BXI:A	32	HHHHHHH	3E0G:A	132	EEEEEEE
LYRRAQG	1IHG:A	309	HHHHHHH	1CI8:A	48	EEEEEEE
LYVKLHN	1I3P:A	276	HHHHHHH	1JIG:A	19	EEEEEEE
METEAVN	1Y10:A	193	HHHHHHH	1Q33:A	221	EEEEEEE
NAIALSA	2QH5:A	91	HHHHHHH	1UCH:A	222	EEEEEEE
NAKTDSI	1VS6:I	42	HHHHHHH	1UYN:X	100	EEEEEEE
NVINTFT	1CJC:A	59	HHHHHHH	3EQZ:A	100	EEEEEEE
PEYLAAF	3C0Y:A	291	HHHHHHH	1NWC:A	341	EEEEEEE
QARAVVL	1IXR:B	161	HHHHHHH	1IV1:A	142	EEEEEEE
QASLLRL	2BEC:A	24	HHHHHHH	1D0K:A	271	EEEEEEE
QAVQAAQ	1SJ7:A	22	HHHHHHH	1APT:E	99	EEEEEEE

Chameleons of Length Seven
Continued

SEQUENCE	PROTEIN	POSITION	SECONDARY	PROTEIN	POSITION	SECONDARY
RAVALRA	3BNI:A	66	HHHHHHH	1JJ2:B	79	EEEEEEE
RFVLALL	1SFK:A	22	HHHHHHH	1QZ8:B	38	EEEEEEE
RGVCTVV	2EJC:A	125	HHHHHHH	2J00:L	33	EEEEEEE
RLERVLE	2FUG:5	1	HHHHHHH	2PWY:A	216	EEEEEEE
RVIVAGL	1X0U:A	495	HHHHHHH	2JA1:A	111	EEEEEEE
SAVVLSA	1GW5:B	252	HHHHHHH	2J7N:A	109	EEEEEEE
SLLDYEV	2RCA:A	223	HHHHHHH	2R5R:A	124	EEEEEEE
SLNSLRF	2REP:A	362	HHHHHHH	1SLQ:A	172	EEEEEEE
SLSVTLQ	2DI3:A	80	HHHHHHH	2HJS:A	246	EEEEEEE
SVTAFLN	1PT6:A	24	HHHHHHH	2A6X:A	123	EEEEEEE
TNALHFV	2VPW:C	16	HHHHHHH	2CWM:A	122	EEEEEEE
TVRENLA	2PEI:A	72	HHHHHHH	2REG:A	81	EEEEEEE
TVSARLF	2P7N:A	113	HHHHHHH	1JU3:A	443	EEEEEEE
VALELYV	1H31:A	238	HHHHHHH	2DQ6:A	208	EEEEEEE
VAQLRIA	1XWY:A	114	HHHHHHH	1YRW:A	144	EEEEEEE
VASLLVK	2RH8:A	20	HHHHHHH	1BZY:C	158	EEEEEEE
VEGRAIL	3F13:A	117	HHHHHHH	3CMB:A	267	EEEEEEE
VGISAVM	1BS2:A	452	HHHHHHH	1TDQ:A	174	EEEEEEE
VGTELNA	1NOF:A	240	HHHHHHH	1VH9:A	82	EEEEEEE
VKTIKMF	2VTY:A	155	HHHHHHH	1PGS:A	39	EEEEEEE
VLDRVES	1GC5:A	257	HHHHHHH	2CW7:A	498	EEEEEEE
VLYVKLH	1I3P:A	275	HHHHHHH	1JIG:A	18	EEEEEEE
VRLAALN	2BKU:B	196	HHHHHHH	2E3V:A	81	EEEEEEE
VSYAAGA	1NGS:A	445	HHHHHHH	3BCZ:A	282	EEEEEEE
VTAFLND	1PT6:A	25	HHHHHHH	2A6X:A	124	EEEEEEE
VVETLAR	1PS6:A	95	HHHHHHH	2Q78:A	76	EEEEEEE
VYERFKA	3E8J:A	164	HHHHHHH	2GJG:A	180	EEEEEEE
YALEGAV	3BQS:A	55	HHHHHHH	1BO5:O	299	EEEEEEE
YLQGIEF	2EBB:A	34	HHHHHHH	1JI6:A	273	EEEEEEE
YVLGIEV	2HP0:A	145	HHHHHHH	2HOE:A	87	EEEEEEE
YVREEVF	1Q2Y:A	16	HHHHHHH	1OHG:A	46	EEEEEEE

Table 14 - Distribution of Chameleon Protein Sequences by CATH Architecture

(Includes 7s which are part of 8s)

1.0	Mainly Alpha	
1.10	Orthogonal Bundle	11
1.20	Up-down Bundle	5
1.25	Alpha Horseshoe	2
1.40	Alpha solenoid	0
1.50	Alpha/alpha barrel	0
2.0	Mainly Beta	
2.10	Ribbon	0
2.20	Single Sheet	0
2.30	Roll	2
2.40	Beta barrel	8
2.50	Clam	0
2.60	Sandwich	11
2.70	Distorted Sandwich	1
2.80	Trefoil	0
2.90	Orthogonal Prism	0
2.100	Aligned Prism	0
2.102	3-layer Sandwich	0
2.105	3 Propellor	0
2.110	4 Propellor	0
2.115	5 Propellor	0
2.120	6 Propellor	0
2.130	7 Propellor	0
2.140	8 Propellor	0
2.150	2 Solenoid	0
2.160	3 Solenoid	0
2.170	Beta Complex	1
3.0	Mixed alpha-beta	
3.10	Roll	3
3.15	Super Roll	0
3.20	Alpha-Beta Barrel	4
3.30	2-Layer Sandwich	16
3.40	3-Layer(aba) Sand.	28
3.45	3-Layer(aab) Sandwich	0
3.50	3-Layer(bba) Sandwich	1
3.55	3-Layer(bab) Sandwich	0
3.60	4-Layer Sandwich	2
3.65	Alpha-beta prism	0
3.70	Box	0
3.75	5-stranded Propellor	0
3.80	Alpha-Beta Horseshoe	2
3.90	Alpha-Beta Complex	4
3.100	RibosomalProtein L15; Chain K; domain2	0
4.0	Few Secondary Structures	
4.1	Irregular	0
	Not assigned	86
	Total	187

3.2.4 Analysis – Shannon's Uncertainty Measure

3.2.4.1 Identifying flanks

Shannon's uncertainty measure was calculated for the central amino acid in all triples for each protein in the chameleon database using CB513 as the reference set. The chameleons of length 7 were identified along with the fourteen amino acids to the N terminus and C terminus of the chameleon regions. The two neighboring regions were then split in half creating a far N terminus flank(A), a N terminus flank(B), a chameleon(C), a C terminus flank(D) and a far C terminus flank(E). Each of these is seven amino acids long.

3.2.4.2 T-Tests

The average uncertainty for each region was computed by sequence and overall. The average uncertainty by position was also computed. The average uncertainty by position is shown in Figure 19.

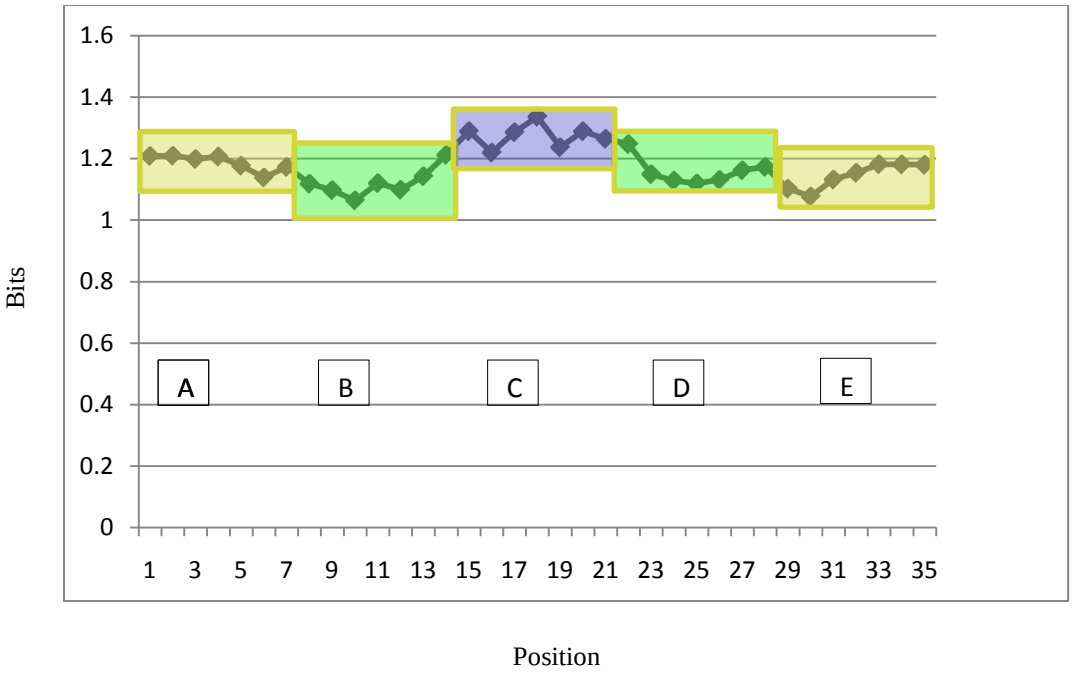


Figure 19 - Average Uncertainty by Position – Chameleons

The hypotheses given earlier were

1-H_0: $U_{cham} = U_{flank}$	2-H_0: $U_{flank} = U_{other}$	3-H_0: $U_{cham} = U_{other}$
1-H_1: $U_{cham} \neq U_{flank}$	2-H_1: $U_{flank} \neq U_{other}$	3-H_1: $U_{cham} \neq U_{other}$

Given the data, these hypotheses can be expressed as:

1a-H_0: $U_C = U_B$	2a-H_0: $U_B = U_A$	3a-H_0: $U_C = U_A$
1a-H_1: $U_C \neq U_B$	2a-H_1: $U_B \neq U_A$	3a-H_1: $U_C \neq U_A$
1b-H_0: $U_C = U_D$	2b-H_0: $U_D = U_E$	3b-H_0: $U_C = U_E$
1b-H_1: $U_C \neq U_D$	2b-H_1: $U_D \neq U_E$	3b-H_1: $U_C \neq U_E$

Note that since there is both a C terminus and N terminus flank and a C terminus and N terminus far flank there are now twice as many hypotheses to test. In addition, there are the following hypotheses which test the C terminus and N terminus flanks against each other:

4a-H_0: $U_B = U_D$	5a-H_0: $U_A = U_D$
4a-H_1: $U_B \neq U_D$	5a-H_1: $U_A \neq U_D$
4b-H_0: $U_A = U_E$	5b-H_0: $U_B = U_E$
4b-H_1: $U_A \neq U_E$	5b-H_1: $U_B \neq U_E$

Student T-tests were conducted on each hypothesis. All of the T-tests were two sample, two tail, unequal variance tests. The results are given below.

	Hypothesis
9.6E-18 =	P value of T-test chameleons vs N terminus flank (1a-H_0)
2.9E-12 =	P value of T-test chameleons vs C terminus flank (1b-H_0)
0.00046 =	P value of T-test N terminus flank vs far N terminus flank (2a-$H_{0)}$
0.44 =	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
6.4E-08 =	P value of T-test chameleon vs far N terminus flank (3a-H_0)
4.3E-13 =	P value of T-test chameleon vs far C terminus flank (3b-H_0)
0.050 =	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.022 =	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.11 =	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.27 =	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 15 - T-Test Results – Chameleons – Uncertainty

3.3.4.3 Bonferroni Correction

When conducting several statistical tests on the same data one needs to be concerned with multiple test error. Levels of statistical significance, (α), are established assuming one test. With an alpha of .05 one would assume that one would wrongly declare a relationship once in 20 cases. If forty tests are run one might expect two errors made simply by chance. To guard against this possibility, one may divide the alpha value by the number of tests. In this way, the family of tests share a risk of alpha. This is called a Bonferroni multiple test correction.

3.2.4.4 Results

The T-test shows that of the ten null hypotheses tested, three (2b,5a,5b) cannot be rejected at the .05 level and two additional ones (4a and 4b) cannot be rejected at the .01 significance level. From this the following conclusions may be drawn:

1. Chameleons on average, have a higher uncertainty than the regions which surround them.
2. The flanking region to the N terminus of the chameleon on average, has a significantly lower uncertainty than the region seven amino acids closer to the N terminus.
3. The flanking regions to the C terminus and to the N terminus chameleon are not significantly different from each other at the .01 level but are significantly different at the .05 level. When corrected for multiple tests the flanking regions are not statistically different.

This analysis supports the idea that chameleons take their secondary structure from their surroundings. The highly significant difference in uncertainty between the far N terminus region and N terminus flanking region support the notion that the flanking regions themselves are special and may be important in enabling a chameleon secondary structure to form.

3.2.5 Analysis – Chou Fasman

In order to determine the degree to which uncertainty is determined solely by the inherent secondary structural tendencies of the individual amino acids themselves, the above analysis was repeated using Chou Fasman numbers in lieu of Shannon's uncertainty calculations. Since there are three sets of Chou Fasman numbers: one for helix, one for beta sheets and one for turns, this analysis was done three times. The results of these analyses follow.

3.2.5.1 Alpha helix numbers

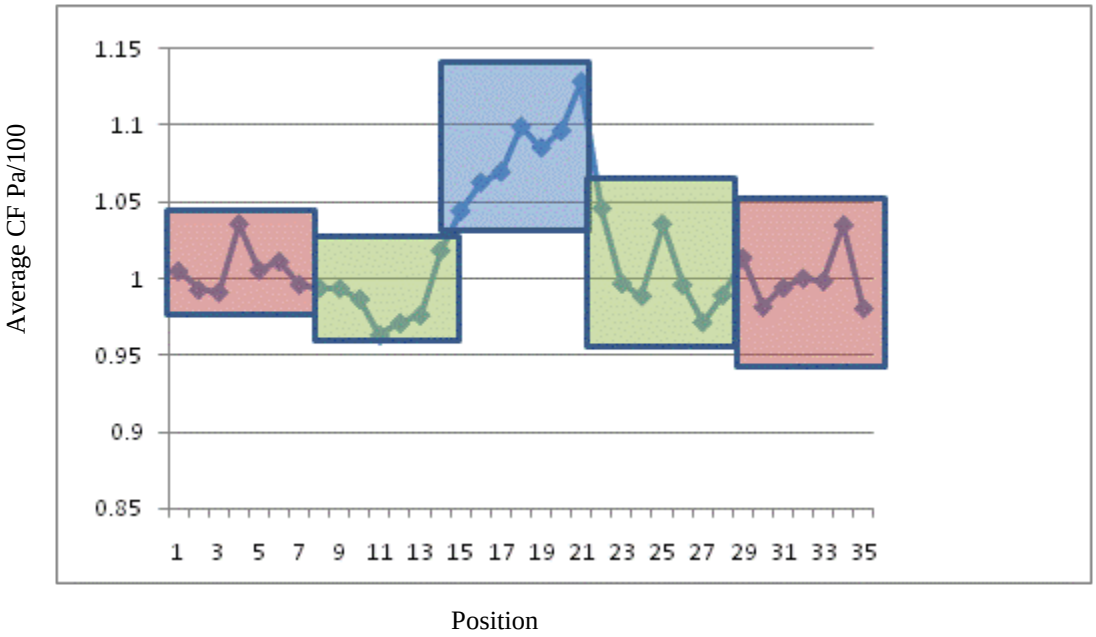


Figure 20 - Average Chou Fasman Pa Number by Position - Chameleons

The same analyses were completed using the Pa numbers. The results were:

	Hypothesis
5.6E-17 =	P value of T-test chameleons vs N terminus flank (1a-H_0)
5.4E-13 =	P value of T-test chameleons vs C terminus flank (1b-H_0)
0.10 =	P value of T-test N terminus flank vs far N terminus flank (2a-$H_{0)}$
0.80 =	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
3.8E-12 =	P value of T-test chameleon vs far N terminus flank (3a-H_0)
3.0E-13 =	P value of T-test chameleon vs far C terminus flank (3b-H_0)
0.13 =	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.68 =	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.87 =	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.22 =	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 16 - T-Test Results – Chameleons – Chou Fasman Pa

As one can see, of the ten hypotheses tested six cannot be rejected at the .05 significance level (2a,2b,4a,4b,5a and 5b). All five of the hypotheses rejected at the .01 significance level using Shannon's information theory are now rejected at the .05 level. In addition, hypothesis 2a which had a probability of 4.6E-4 using information theory cannot be rejected at the .05 level using Chou Fasman's Pa numbers. Only the hypotheses comparing the chameleons to their flanking and far flanking regions can be rejected following multiple test correction.

3.2.5.2 Beta sheet numbers

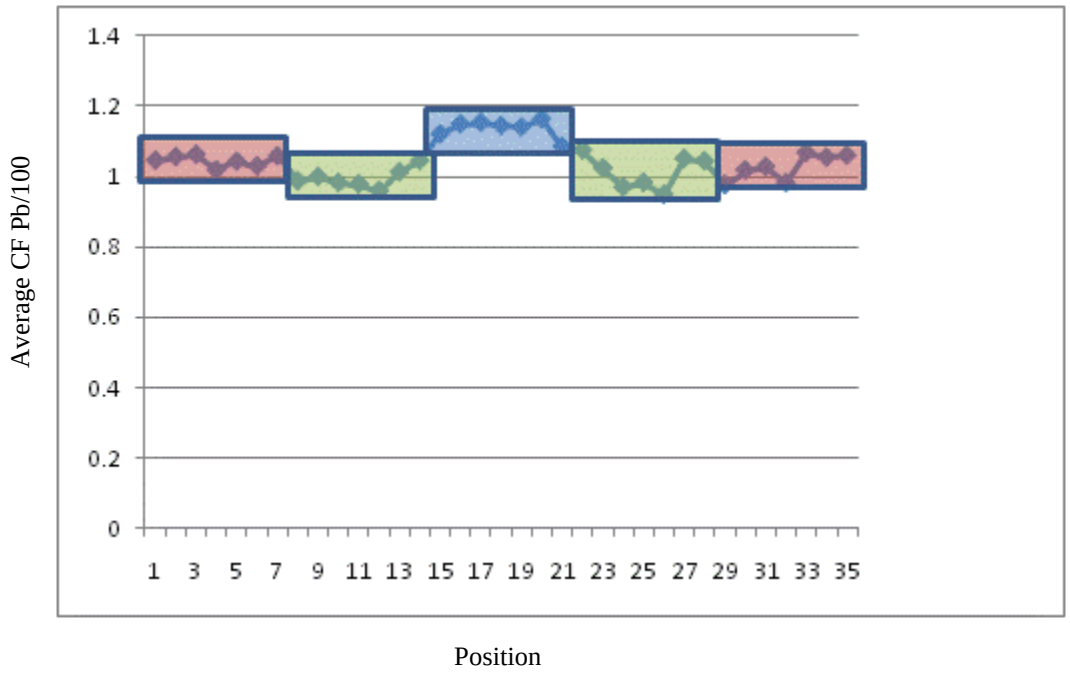


Figure 21 - Average Chou Fasman Pb Number by Position - Chameleons

The same analyses were completed using the Pb numbers. The results are shown in Table 17.

		Hypothesis
4.0E-25	=	P value of T-test chameleons vs N terminus flank (1a-H_0)
1.3E-19	=	P value of T-test chameleons vs C terminus flank (1b-H_0)
0.00027	=	P value of T-test N terminus flank vs far N terminus flank (2a-H_0)
0.34	=	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
1.6E-11	=	P value of T-test chameleon vs far N terminus flank (3a-H_0)
1.2E-15	=	P value of T-test chameleon vs far C terminus flank (3b-H_0)
0.16	=	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.19	=	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.023	=	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.020	=	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 17 - T-Test Results – Chameleons – Chou Fasman Pb

The T-test shows that of the ten null hypotheses tested, three (2b,4a,4b) cannot be rejected at the .05 level and two additional ones (5a and 5b) cannot be rejected at the .01 significance level. Like the Pa analyses the four hypotheses involving chameleons can be rejected. In addition the hypothesis comparing the near N flank to the far N flank can also be rejected even after correction for multiple testing. The others cannot.

3.2.5.3 Beta Turn numbers

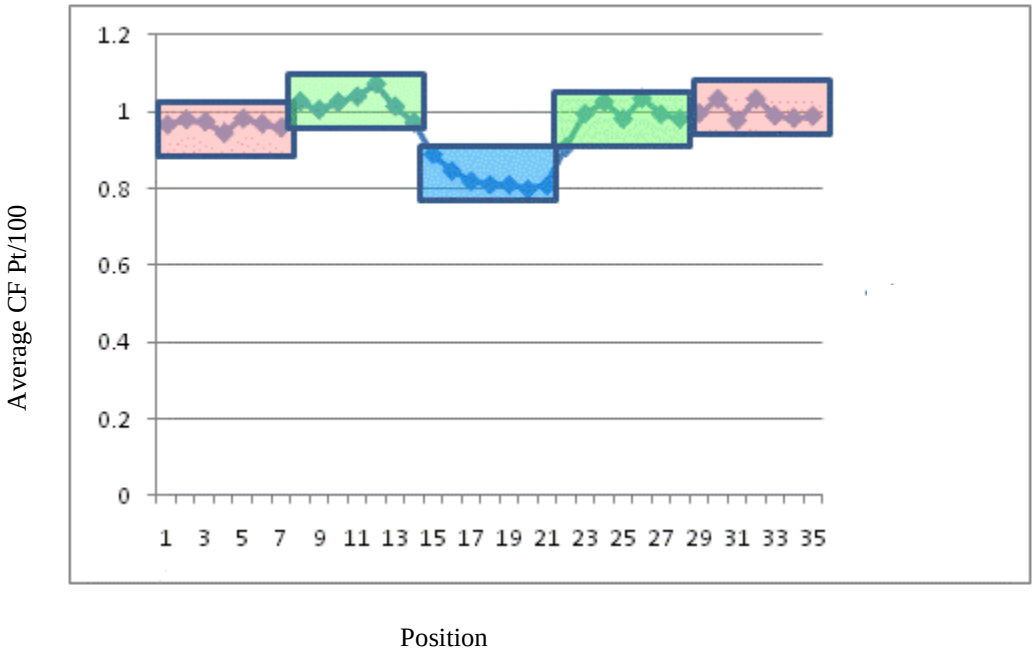


Figure 22 - Average Chou Fasman Pt Number by Position - Chameleons

The same analyses were completed using the Pt numbers. The results were:

Hypothesis

1.6E-37 = P value of T-test chameleons vs N terminus flank (1a-H_0)
1.2E-27 = P value of T-test chameleons vs C terminus flank (1b-H_0)

0.00046 = P value of T-test N terminus flank vs far N terminus flank (2a-H_0)
0.19 = P value of T-test C terminus flank vs far C terminus flank (2b-H_0)

2.4E-21 = P value of T-test chameleon vs far N terminus flank (3a-H_0)
6.8E-27 = P value of T-test chameleon vs far C terminus flank (3b-H_0)

0.025 = P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.31 = P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)

0.19 = P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.0094 = P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 18 - T-Test Results – Chameleons – Chou Fasman Pt

The T-test shows that of the ten null hypotheses tested, three (2b,4b,5b) cannot be rejected at the .05 level and one (4a) cannot be rejected at the .01 significance level.

3.2.6 Comparison of Information Uncertainty to Chou Fasman Results

Hypothesis	Description	Prob Uncertainty	Prob Pa	Prob Pb	Prob Pt
1a-Ho	Chameleon vs N Terminus Flank	9.6E-18	5.6E-17	4E-25	1.6E-37
1b-Ho	Chameleon vs C Terminus Flank	2.9E-12	5.4E-13	1.3E-19	1.2E-27
2a-Ho	N Terminus Flank vs Far N Term. Flank	0.00046	0.1	0.00027	0.00046
2b-Ho	C Terminus Flank vs Far C Term. Flank	0.44	0.8	0.34	0.19
3a-Ho	Chameleon vs Far N Terminus Flank	6.4E-08	3.8E-12	1.6E-11	2.4E-21
3b-Ho	Chameleon vs Far C Terminus Flank	4.3E-13	3E-13	1.2E-15	6.8E-27
4a-Ho	N Terminus Flank vs C Terminus Flank	0.05	0.13	0.16	0.025
4b-Ho	Far N Term. Flank vs Far C Term. Flank	0.022	0.68	0.19	0.31
5a-Ho	Far N Term.Flank vs C Term. Flank	0.11	0.87	0.023	0.19
5b-Ho	N Term. Flank vs Far C Term. Flank	0.27	0.22	0.02	0.0094
	Significant with Bonferroni correction				
	Insignificant with Bonferroni correction				

Table 19 - Uncertainty vs Chou-Fasman Results – Chameleon

The table above highlights those hypotheses which are rejected at the .05 significance level. It shows that there is consensus among the methods on five hypotheses (1a,1b,2b,3a,3b). With Bonferroni correction there is consensus on four more (4a,4b,5a,5b). On one hypothesis (2a), three agree (uncertainty, Pb, and Pt).

3.2.7 Interpretation of Results – Chameleons

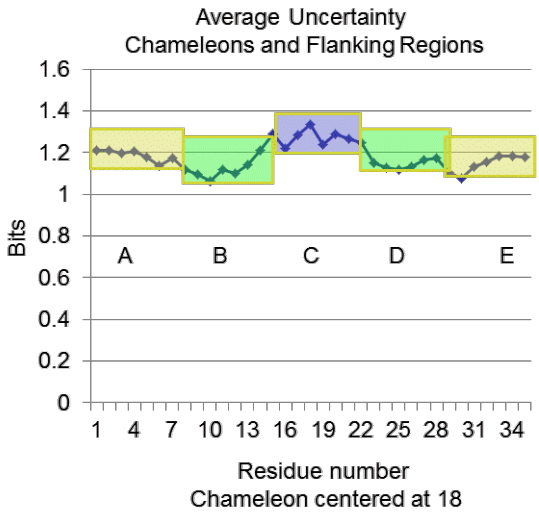


Figure 23 - Chameleon – Uncertainty

If one looks at the graph of the uncertainty measure one sees an average of 1.2 bits in the far N terminus flank followed by a decline to 1.1 in the measure as one moves toward the chameleon. The measure then rises to a peak of 1.33 followed by a decline to 1.08 before returning to an average of 1.18 in the far C flank.

While the growth in the measure, peaking in the chameleon, is clear, an increase of 20% from the low and a 10% increase from the average, it is not unexpected. The more

interesting phenomenon is the 8-10% decline in uncertainty in the flanking regions prior and after the chameleon.

It is believed that the chameleon takes its structure from those around it. Tankano *et al*. called this conformational contagion (Section 2.3.10). If this is the case we would expect a decline in uncertainty due to ordered conditions which would favor one structure over another. One might also expect that the amino acids immediately next to the chameleon would have the same secondary structure. To test this idea the neighboring amino acids were counted.

Same as chameleon	277	81.50%	
Different than chameleon	63	18.50%	
Coil	63		
Helix	0		
Sheet	0		
Cases with both neighbors coil			11
sheet chameleons			11
helix chameleons			0
Cases where one neighbor is coil			
sheet chameleons			28
helix chameleons			13
Total Different		(11*2) + 28 + 13 =	63

Table 20 - Neighboring Amino Acid Data

3.3 Application - Protein Hinges

3.3.1 Hinge Atlas

Protein hinges are flexible regions connecting two rigid domains (Figure 16). They are often found near or at protein active sites and are therefore an area of growing interest within the protein science community [Flores *et al*., 2007]. As a second test of the method, protein hinges were investigated.

Flores *et al*. have developed two databases of protein hinges which they have made available at the Gerstein lab website. The first is the nonredundant Hinge Atlas. The second is the hinge atlas gold standard. Both data sets are hand annotated collections of hinges in various proteins.

The gold standard has 20 proteins in it. The non redundant set has 220. More importantly, the non redundant set has the amino acid sequence listed for each protein. As a result, the non-redundant set was used. Of these proteins, 202 were two amino acid hinges, 12 involved 3 amino acids, 4 involved 4 amino acids and two had 5. Only the two amino acid hinges were retained. In reviewing the remaining set, some proteins with labels like A, test, model2 and www were clearly not intended to be used generally. These were removed. Only those protein hinges with PDB codes and full secondary structure were retained. As before, five regions were identified for each hinge: the hinge region (hinge -2 amino acids – hinge +2 amino acids); the six amino acids directly toward the C terminus and N terminus of the hinge region (flanks) and the six amino acids to the N terminus and C terminus of the flanking regions (far flanks). Unlike the chameleon analysis, all regions are six instead of seven amino acids long. Any hinges which were

within 14 residues of the ends of the protein were discarded. Any hinges which had regions which overlapped the regions associated with any other hinge were also discarded to avoid confounding the data. The number of proteins remaining was 46 containing 65 hinges. The uncertainty for each central position was then computed using CB513 as the reference set.

Table 21 - Selected 2-Residue Protein Hinges

Protein	N terminus residue	C terminus residue
13pkD	184	185
172lA	129	130
1af7A	33	34
1alsA	87	88
1amcA	28	29
1be3E	66	67
1bj6A	21	22
1bmtA	92	93
1bsra	19	20
1c0aA	111	112
1c0aA	270	271
1c0aA	318	319
1c0aA	350	351
1c0aA	421	422
1cg3A	224	225
1cgjE	195	196
1cwuB	51	52
1d9nA	20	21
1dotA	92	93
1dotA	246	247
1dotA	431	432
1dpeA	262	263
1dpeA	478	479
1dr8A	253	254
1eiaA	130	131
1ex6A	30	31
1fdmA	22	23
1fqbA	314	315
1hrdC	19	20
1hrdC	89	90
1hrdC	205	206
1hrdC	305	306

Protein	N terminus residue	C terminus residue
1hreA	28	29
1hup_1	29	30
1iskB	42	43
1iskB	103	104
1iwoA	60	61
1iwoA	115	116
1iwoA	244	245
1jejA	170	171
1jfjA	64	65
1l5bA	50	51
1lila1	106	107
1oibB	255	256
1oibB	21	22
1osa	42	43
1pbnA	254	255
1rckA	21	22
1rkmA	485	486
1rkmA	270	271
1roda	54	55
1tdeA	114	115
1tdeA	245	246
1vkxA	170	171
1vpe	186	187
1vpe	369	370
1zxq_1	87	88
2ctsA	275	276
2gvaB	26	27
2paiA	16	17
3bjlB	110	111
3lip	20	21
9ldta2	206	207
9ldta2	238	239
9ldta2	280	281

Table 22 - Distribution of Selected Protein Hinges Sequences by CATH Architecture

1.0	Mainly Alpha	
1.10	Orthogonal Bundle	7
1.20	Up-down Bundle	1
1.25	Alpha Horseshoe	0
1.40	Alpha solenoid	0
1.50	Alpha/alpha barrel	0
2.0	Mainly Beta	
2.10	Ribbon	1
2.20	Single Sheet	0
2.30	Roll	1
2.40	Beta barrel	5
2.50	Clam	0
2.60	Sandwich	4
2.70	Distorted Sandwich	0
2.80	Trefoil	0
2.90	Orthogonal Prism	0
2.100	Aligned Prism	0
2.102	3-layer Sandwich	0
2.105	3 Propellor	0
2.110	4 Propellor	0
2.115	5 Propellor	0
2.120	6 Propellor	0
2.130	7 Propellor	0
2.140	8 Propellor	0
2.150	2 Solenoid	0
2.160	3 Solenoid	0
2.170	Beta Complex	1
3.0	Mixed alpha-beta	
3.10	Roll	8
3.15	Super Roll	0
3.20	Alpha-Beta Barrel	0
3.30	2-Layer Sandwich	5
3.40	3-Layer(aba) Sand.	20
3.45	3-Layer(aab) Sandwich	0
3.50	3-Layer(bba) Sandwich	2
3.55	3-Layer(bab) Sandwich	0
3.60	4-Layer Sandwich	0
3.65	Alpha-beta prism	0
3.70	Box	0
3.75	5-stranded Propellor	0
3.80	Alpha-Beta Horseshoe	2
3.90	Alpha-Beta Complex	3
3.100	RibosomalProtein L15; Chain K; domain2	0
4.0	Few Secondary Structures	
4.1	Irregular	2
	Not assigned	6
	Total	66

3.3.2 Hypotheses

The question to be investigated was, “Do hinge regions exhibit greater flexibility than their flanking regions (rigid domains), as measured by this method?” The same hypotheses with the hinge region replacing the chameleon region were posed and tested. To wit:

1-H_0: $U_{hinge} = U_{flank}$	2-H_0: $U_{flank} = U_{other}$	3-H_0: $U_{hinge} = U_{other}$
1-H_1: $U_{hinge} \neq U_{flank}$	2-H_1: $U_{flank} \neq U_{other}$	3-H_1: $U_{hinge} \neq U_{other}$

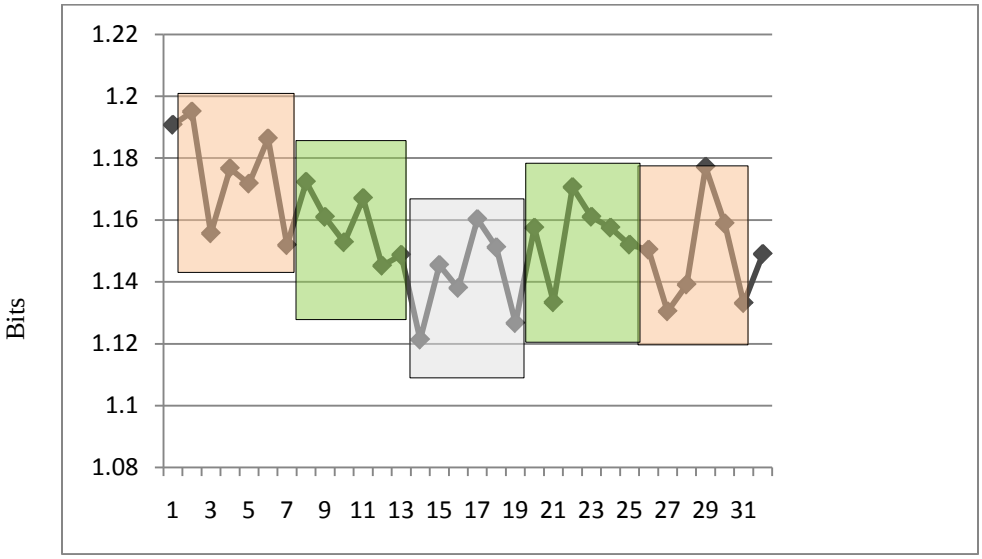


Figure 24 - Average Uncertainty by Position - Hinges

As before, these hypotheses can be expressed as:

1a-H_0: $U_C = U_B$	2a-H_0: $U_B = U_A$	3a-H_0: $U_C = U_A$
1a-H_1: $U_C \neq U_B$	2a-H_1: $U_B \neq U_A$	3a-H_1: $U_C \neq U_A$
1b-H_0: $U_C = U_D$	2b-H_0: $U_D = U_E$	3b-H_0: $U_C = U_E$
1b-H_1: $U_C \neq U_D$	2b-H_1: $U_D \neq U_E$	3b-H_1: $U_C \neq U_E$

4a-H_0: $U_B = U_D$ 5a-H_0: $U_A = U_D$
4a-H_1: $U_B \neq U_D$ 5a-H_1: $U_A \neq U_D$

4b-H_0: $U_A = U_E$ 5b-H_0: $U_B = U_E$
4b-H_1: $U_A \neq U_E$ 5b-H_1: $U_B \neq U_E$

3.3.3 Analysis – Shannon's Uncertainty Measure

There were 65 hinges remaining in the dataset. Student T-tests were conducted on each hypothesis. All of the T-tests were two sample, two tail, unequal variance tests using Excel's built-in function. The results are given below.

		Hypothesis
0.87	=	P value of T-test hinge vs N terminus flank (1a-H_0)
0.43	=	P value of T-test hinge vs C terminus flank (1b-H_0)
0.079	=	P value of T-test N terminus flank vs far N terminus flank (2a-$H_{0)}$
0.28	=	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
0.041	=	P value of T-test hinge vs far N terminus flank (3a-H_0)
0.79	=	P value of T-test hinge vs far C terminus flank (3b-H_0)
0.58	=	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.021	=	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.14	=	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.68	=	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 23 - T-Test Results Hinges - Uncertainty

3.3.4 Results

As one can see, the only null hypotheses which can be rejected based on these results at the .05 significance level are 3a: hinge vs far N terminus flank and 4b: far N terminus flank vs far C terminus flank. All the others must be accepted based on these tests.

When corrected for multiple tests these two hypotheses also become statistically insignificant.

3.3.5 Analysis – Chou Fasman

As before, the above analysis was repeated using Chou Fasman numbers instead of Shannon's uncertainty calculations. The results of these analyses follow.

3.3.5.1 Alpha helix numbers

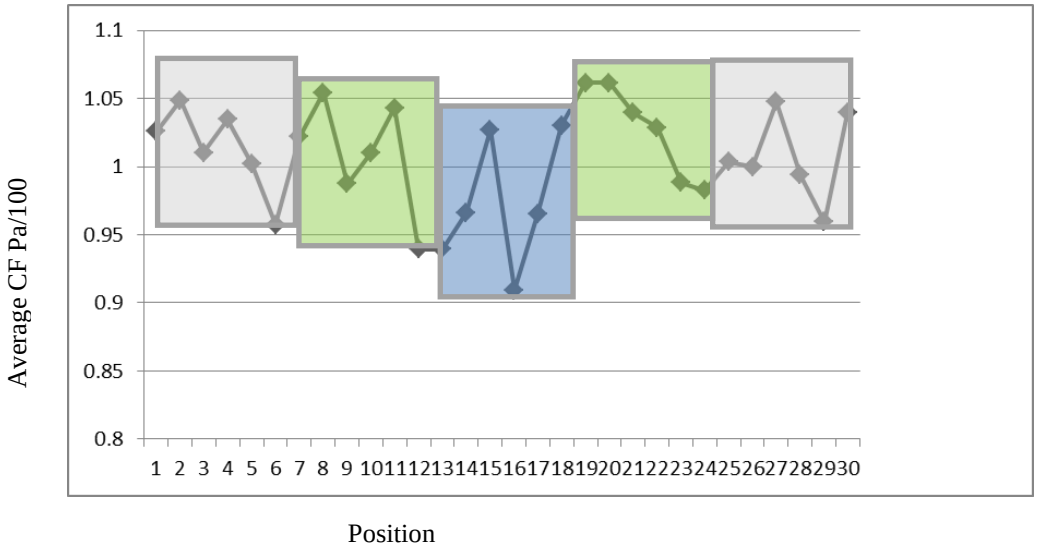


Figure 25 - Average Chou Fasman Pa Number by Position - Hinges

The same analyses were completed using the Pa numbers. The results were:

		Hypothesis
0.44	=	P value of T-test hinge vs N terminus flank (1a-H_0)
0.50	=	P value of T-test hinge vs C terminus flank (1b-H_0)
0.85	=	P value of T-test N terminus flank vs far N terminus flank (2a-$H_{0)}$
0.51	=	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
0.34	=	P value of T-test hinge vs far N terminus flank (3a-H_0)
0.98	=	P value of T-test hinge vs far C terminus flank (3b-H_0)
0.91	=	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.34	=	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.76	=	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.45	=	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 24 - T-Test Results – Hinges – Chou Fasman Pa

As one can see, of the ten hypotheses tested none can be rejected at the .05 significance using Chou Fasman's Pa numbers.

3.3.5.2 Beta sheet numbers

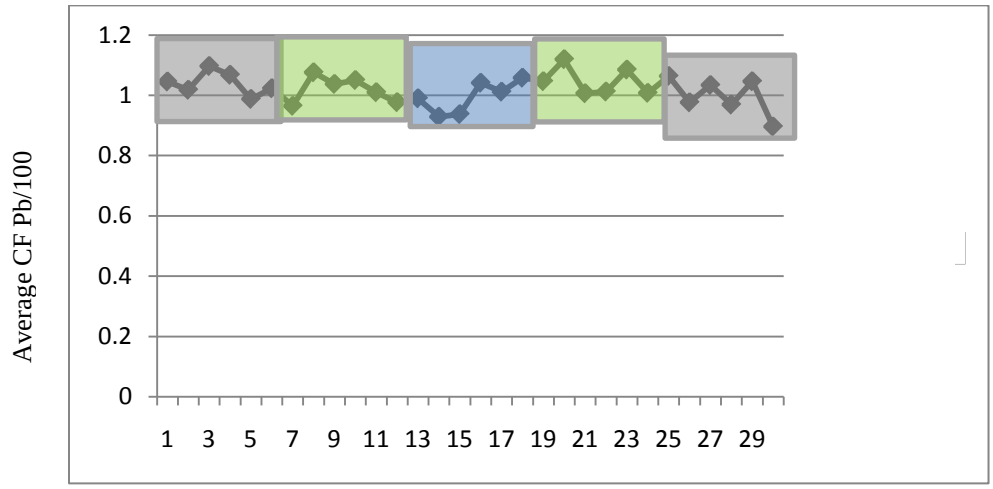


Figure 26 - Average Chou Fasman Pb Number by Position – Hinges

The same analyses were completed using the Pb numbers. The results were:

		Hypothesis
0.36	=	P value of T-test hinge vs N terminus flank (1a-H_0)
0.067	=	P value of T-test hinge vs C terminus flank (1b-H_0)
0.43	=	P value of T-test N terminus flank vs far N terminus flank (2a-H_0)
0.077	=	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
0.10	=	P value of T-test hinge vs far N terminus flank (3a-H_0)
0.91	=	P value of T-test hinge vs far C terminus flank (3b-H_0)
0.31	=	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.12	=	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.81	=	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.41	=	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 25 - T-Test Results – Hinges – Chou Fasman Pb

The T-test shows that of the ten null hypotheses none can be rejected at the .05 level.

3.3.5.3 Beta turn numbers

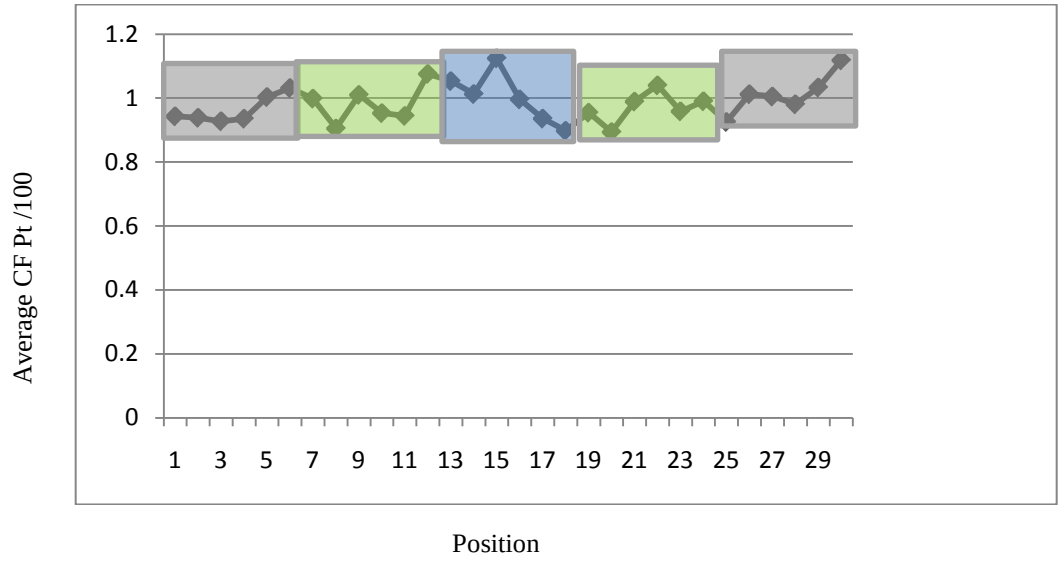


Figure 27 - Average Chou Fasman Pt Number by Position - Hinges

The same analyses were completed using the Pt numbers. The results were

		Hypothesis
0.42	=	P value of T-test hinge vs N terminus flank (1a-H_0)
0.26	=	P value of T-test hinge vs C terminus flank (1b-H_0)
0.50	=	P value of T-test N terminus flank vs far N terminus flank (2a-H_0)
0.13	=	P value of T-test C terminus flank vs far C terminus flank (2b-H_0)
0.14	=	P value of T-test hinge vs far N terminus flank (3a-H_0)
0.74	=	P value of T-test hinge vs far C terminus flank (3b-H_0)
0.71	=	P value of T-test N terminus flank vs C terminus flank (4a-H_0)
0.062	=	P value of T-test far N terminus flank vs far C terminus flank (4b-H_0)
0.78	=	P value of T-test far N terminus flank vs C terminus flank (5a-H_0)
0.24	=	P value of T-test N terminus flank vs far C terminus flank (5b-H_0)

Table 26 - T-Test Results – Hinges – Chou Fasman Pt

The T-test shows that of the ten null hypotheses tested, none can be rejected at the .05 level.

3.3.6 Comparison of Information Uncertainty to Chou Fasman Results

Hypothesis	Description	P-value Uncertainty	P-value Pa	P-value Pb	P-value Pt
1a-Ho	Hinge vs N Terminus Flank	0.87	0.44	0.36	0.42
1b-Ho	Hinge vs C Terminus Flank	0.43	0.50	0.067	0.26
2a-Ho	N Terminus Flank vs Far N Term. Flank	0.079	0.85	0.43	0.50
2b-Ho	C Terminus Flank vs Far C Term. Flank	0.28	0.51	0.077	0.13
3a-Ho	Hinge vs Far N Terminus Flank	0.041	0.34	0.1	0.14
3b-Ho	Hinge vs Far C Terminus Flank	0.79	0.98	0.91	0.74
4a-Ho	N Terminus Flank vs C Terminus Flank	0.57	0.91	0.31	0.71
4b-Ho	Far N Term. Flank vs Far C Term. Flank	0.021	0.34	0.12	0.062
5a-Ho	Far N Term.Flank vs C Term. Flank	0.14	0.76	0.81	0.78
5b-Ho	N Term. Flank vs Far C Term. Flank	0.68	0.45	0.4	0.24
		Significant with Bonferroni correction			
		Insignificant with Bonferroni correction			

Table 27 - Uncertainty vs Chou-Fasman Results - Protein Hinges

The table shows that none of the Chou Fasman analyses show a statistical difference. The only statistical differences identified by the uncertainty analysis disappear when corrected for multiple testing.

3.3.7 The interpretation of results – Hinges

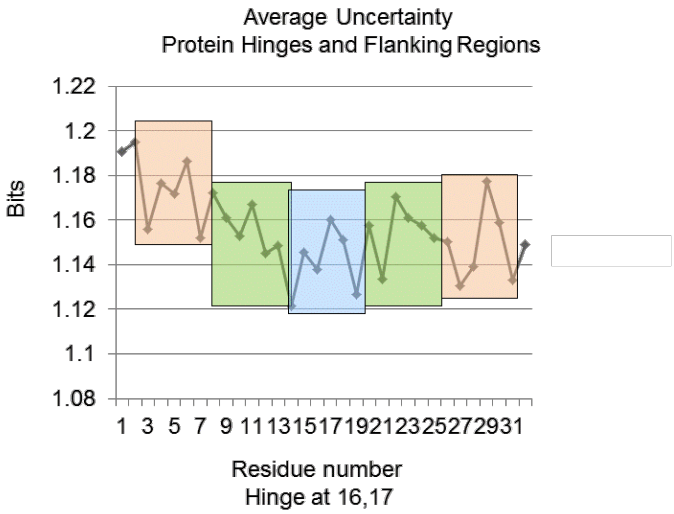


Figure 28 - Average Uncertainty by Position – Hinges

If one looks at the hinge uncertainty graph it starts at 1.19 bits and trends slowly downward amid a lot of variability eventually hitting a low of 1.12 bits at the hinge. It then grows to a high of 1.17 bits before ending at 1.13 bits. This too has a lot of variability.

If one uses the eight to three reduction of H, G -> H; E, B -> E; all others -> C; on the Flores *et al.* data, one calculates that hinges occur 67% of the time in coil, 21% of the time in sheet and 12% helix. Based on Flores *et al*., I assumed that an increase in coil accounted for the lowering of the uncertainty value moving from the far N flank to the hinge. To test this assumption the following analysis was completed.

The amino acids with helix, sheet and coil secondary structure were counted. Since all regions were 6 amino acids long there was no need to average.

	Far NF	NF	Hinge	CF	Far CF	Total
Helix	172	121	82	133	155	663
Sheet	95	94	110	94	60	453
Coil	123	175	198	163	175	834
Total	390	390	390	390	390	1950

Table 28 - Hinge Secondary Structure Counts by Region

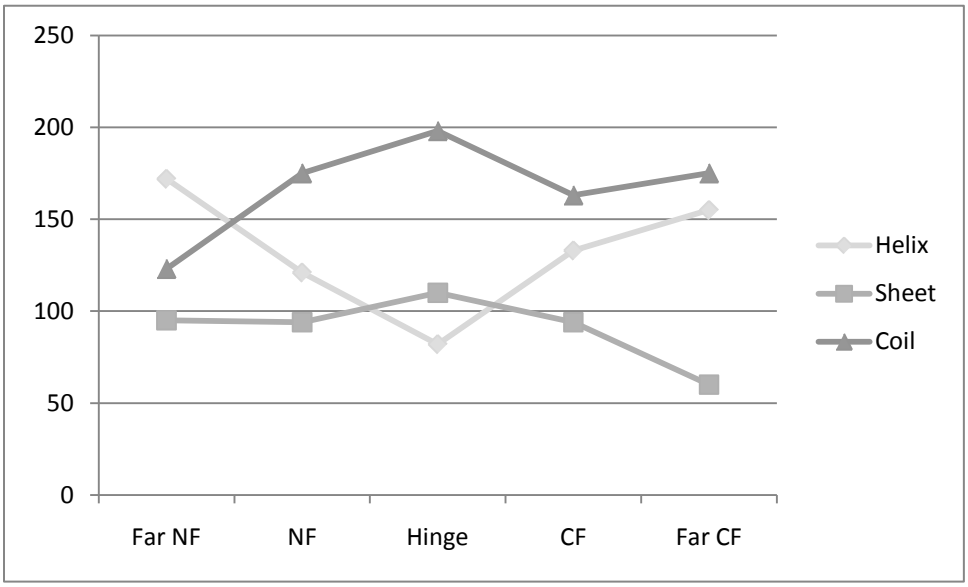


Figure 29 - Hinge Secondary Structure Counts by Region

Applying Shannon's information directly to the values from Table 28 we get the following:

Far NF	NF	Hinge	CF	Far CF	Average
1.54	1.54	1.48	1.55	1.46	1.51

Table 29 - Autoreferenced Information Entropy by Hinge Region (Bits)

We see apparent differences between the far N flank and the hinge and the far N flank and the far CF regions here also, but nothing that would challenge the conclusion from earlier statistics. No statistically significant differences exist among these regions.

3.4 Comparison of work to Kuzenetsov and Rackovsky

Kuzenetsov and Rackovsky 2003 was in part, an inspiration for my work. However, there are significant differences. 1) They measure the dihedral angles of the central residue and compute a propensity to participate in helix, sheet or coil; I compute the propensity to participate within a reference set (CB513) and then look up the appropriate

number for each tripeptide. 2) They group the flanking amino acids (x+/- 1) into three groups based on physio-chemical properties; I use the entire tripeptide as an index to the lookup table. 3) They define five classes of structurally ambivalent peptides (SAPs) and use their measure to investigate the classes, predicting membership and identifying different characteristics based on their (GLP); I developed a large database of only the rarest SAPs (helix-extended) which I investigate and also apply my measure to protein hinges which Kuznetsov and Rackovsky did not consider. 4) Each of their computations are normalized to a global average so that positive values have less uncertainty and negative values have more uncertainty then average. All of my calculations are positive and have values between 0 and $-\log_2(1/x)$ where X is the number of possible outcomes. For x = 3, (helix, extended, coil) it equals approximately 1.585. A 0 corresponds to no uncertainty, 1.585 is the maximum uncertainty for a three state case. (1 is the maximum for a two state case; 2 for a four state case.) 5) Kuznetsov and Rackovsky's structurally ambivalent peptides are all k-mers of length five and six. My analyses were done using only Helix-Extended chameleons of length seven. I also identified and verified nine Helix - Extended chameleons of length eight. The greatest number previously found was two. [Guo *et al*. 2007]

4.0. CONTRIBUTIONS AND FUTURE WORK

4.1 Contributions

I have made four contributions in this work: 1) developed a method for measuring uncertainty in protein sequences; 2) developed a new chameleon database; 3) provided additional support to the idea of conformation contagion in chameleons; and 4) conducted information entropy measurements on protein hinges.

4.1.1 Method for measuring uncertainty

First and foremost, I have developed a method for measuring uncertainty in protein sequences. This method uses an external reference set, Cuff and Barton 513 and Shannon's information theory to compute the information entropy associated with any arbitrary protein sequence. This method was used to measure the information entropy of chameleon sequences and their flanking regions. This measurement showed that "typical" sequences have approximately 1.17 bits and that chameleon sequences have approximately 1.27 bits. The difference is approximately 0.1 bits.

This analysis was repeated using Chou-Fasman numbers (Pa,Pb,Pt) with similar results.

4.1.2 New chameleon database

I also developed a new database of long helix-extended chameleon sequences. Each sequence was carefully reviewed and checked against several criteria. This resulted in identifying nine chameleons of length eight and eighty-five of length seven. The largest collection reported in the literature is two and sixty-three respectively [Guo, 2007].

4.1.3 Support for Conformation Contagion

I applied my metric to the new chameleon database and found that the information-entropy declined in the regions flanking the chameleon. This is explained as area of increased order surrounding the chameleon. An additional analysis was accomplished to check this assumption and it confirmed the assumption. Both analyses provide support for the idea that chameleons take their secondary structure from local sequence interactions. This is termed conformation contagion.

4.1.4 Protein Hinges

My information-entropy metric was also applied to a set of protein hinges. The metric appeared to find two marginally significant relationships which disappeared following Bonferroni multiple test correction. The three Chou Fasman analyses also found no statistically significant relationships among the hinge related regions. However, the average information-entropy across all of the hinge regions (hinge, near and far flanks) using my method and CB513 as a reference was 1.14 bits. The same number using the data as an autoreference was 1.51.

4.2 Future Work

4.2.1 Develop reference set rules

One of the key features of this work is its use of an external reference set (CB513). While it was quite useful in developing and demonstrating the technique, CB513 is not yet suitable to be used as a standard. For example, it does not contain 330 of the 8000 triples possible when using a moving window of size three. While this did not affect the results of the chameleon analysis (less than one half of a percent of chameleon triples

were missing) it could affect other analyses. As part of the development of a standard reference set this must be addressed.

A more general issue is the development of reference set rules. The use of an external reference set is ideal for single sequences but may distort database relationships. The large difference between the CB513 mediated information-entropy for the hinge data 1.14 bits and the autoreferenced value of 1.51bits illustrate the potential for differences. An external reference necessarily produces a mapping of one set of relationships onto another. These projections can both illuminate and distort the underlying relationships.

In order for this technique to gain wide acceptance the strengths and weaknesses of different database configurations need to be explored and rules developed for their proper construction and use. It is possible that a family of database templates can be developed to accomplish specific tasks.

4.2.2 Spatial proximity

The second area is to investigate how the results would change if spatial proximity was used to calculate the uncertainty numbers instead of sequential proximity. It is known that spatial relationships determine secondary structure and proteins often fold so as to put residues which are quite distant in the sequence close together in space. It would be interesting to see how large the effect of this might be on the measurement.

4.2.3 Compare to Kuzenetsov and Rackovsky

The third area is to compare the measurements derived using this method directly to measurements using the Kuzenetsov and Rackovsky method. Kuzenetsov and Rackovsky also measured uncertainty of proteins using information theory. They used triples but rather than a reference set of secondary structures, they used dihedral angles. It would be interesting to compare the two methods directly against a common set of sequences.

4.2.4 Uncertainty vs Function

A fourth area to explore in future work is to see if there is any relationship between a protein's uncertainty score and its function. Are proteins which act as transportation (e.g. hemoglobin) different in their scores than those that provide structure? Are active sites different from non-active sites? The comparison of the information entropy associated with a protein's structure and its function may prove illuminating.

REFERENCES

Anfinsen, C. B. *et al*. "The Kinetics of Formation of Native Ribonuclease During the Reductiion of the Reduced Polypeptide Chain." Proceedings of the National Academy of Science 47.9 (1961): 1309.

Anfinsen, Christian B. "Studies on the Principles that Govern the Folding of Protein Chains." Nobel Lecture. 11 December 1972. 104.

Arndt, C. Information Measures: information and its description in science and engineering. New York: Springer,Berlin, 2001.

Baldi, Pierre *et al*. "Exploiting the Past and the Future in Protein Secondary Structure Prediction." Bioinformatics 15.11 (1999): 937-946.

Branden, Carl and John Tooze. Introduction to Protein Structure. 2nd Edition. New York: Garland Publishing Inc, 1999.

Brownlee, Christen. "The Protein Papers." Proceedings of the National Academy of Science. 2006 <http://www.pnas.org/classics1.shtml>.

Bystroff, Christopher *et al*. "HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins." Journal of Molecular Biology 301 (2000): 173-190.

Campbell, Neil A. *et al*. Biology. 5th Edition. Menlo Park : Benjamin/Cummings, 1999.

Chou, Peter Y and Gerald D Fasman. "Conformational Parameters for Amino Acids in Helical, b-Sheet, and Random Coil Regions Calculated from Proteins.”." Biochemistry 13.2 (1974a): 211-222.

—. "Prediction of Protein Conformation." Biochemistry 13. 2 (1974b): 222-245.

—. "Prediction of Protein Secondary Structure." Advances in Enzymology . Vol. 47. Ed. Alton Meister. New York : John Wiley and Sons , 1978. 45-148.

Cohen, B I, S R Presnell and F E Cohen. "Origins of Structural Diversity Within Sequentially Identical Hexapeptides." Protein Science 2 (1993): 2134-2145.

Cover, Thomas and Joy Thomas. Elements of Information Theory. New York : John Wiley and Sons , 1991.

Crooks, Gavin E and Steven E Benner. "Protein Secondary Structure: Entropy, Correlations and Prediction." Bioinformatics 20.10 (2004): 1603-1611.

Cuff, James A and Geoffery J Barton. "Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction." PROTEINS: Structure, Function and Genetics 34 (1999): 508-519.

Ding, Young Sheng *et al*. "Using Maximum Entropy Model to Predict Protein Secondary Structure with a Single Sequence." Protein & Peptide Letters 16 (2009): 552-560.

Dor, Ofer and Yaoqui Zhou. "Achieving 80% Ten Fold Cross-validated Accuracy for Secondary Structure Prediction by Large Scale Training." PROTEINS: Structure, Function, and Bioinformatics 66 (2007): 838-845.

Drenth, Jan. Principles of Protein X-Ray Crystallography. 3rd Edition. New York: Springer, 2007.

Emekli, Ugur *et al*. "HingeProt: automated prediction of hinges in protein structures." Proteins 70 (2008): 1219-1227.

Fersht, Alan. Structure and Mechanism in Protein Science. NewYork: W H. Freeman and Company, 1999.

Flores, Samuel *et al*. "Hinge Atlas: Relating Protein Sequence to Sites of Structural Flexibility." BMC Bioinformatics 8.167 (2007).

Garnier, J, Osguthorpre D J and Robson B. "Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins." Journal of Molecular Biology 120 (1978): 97-120.

Garret, Reginald H and Charles M Grisham. Biochemistry. 3rd Edition. Belmont CA: Brooks Cole, 2005.

Guo, Jun-Tao *et al*. "Analysis of Chameleon Sequences and Their Implications in Biological Processes." Proteins: Structure, Function, and Bioinformatics 67 (2007): 548-558.

Hartley, Ralph V. "Transmission of Information." Bell System Technical Journal (July 1928): 535–563.

Hu, Hae-Jin *et al*. "Improved Protein Secondary Structure Prediction Using Support Vector Machine With a New Encoding Scheme and an Advanced Tertiary Classifier." IEEE Transactions on Nanobioscience 3.4 (2004): 265-271.

Jacoboni, Irene *et al*. "Predictions of Protein Segments with the Same Aminoacid Sequemce and Different Secondary Structure: A Benchmark for Predictive Methods." Proteins: Structure, Function, and Genetics 41 (2000): 535-544.

Jones, David T. "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices." Journal of Molecular Biology 292 (1999): 195-202.

Kabsch, Wofgang and Christian Sander. "Dictionary Of Protein Secondary Structure: Pattern Recognition Of Hydrogen-Bonded And Geometrical Features." Biopolymers 22.12 (1983): 2577-2637.

—. "On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations." Proceedings of the National Academy of Science USA (1984): 1075-1078.

Katzman, Sol *et al*. "Predict-2nd: a tool for generalized protein local structure prediction." Bioinformatics 24.21 (2008): 2453-2459.

Keating, Kevin *et al*. "StoneHinge: Hinge prediction by network analysis of individual protein structures." Protein Science 18 (2009): 359-371.

Kim, Hyunsoo and Haesun Park. "Protein Secondary Structure Prediction Based on an Improved Support Vector Machines Approach." Protein Engineering 16.8 (2003): 533-560.

Kimball, John W. "Hydrogen Bonds." BiologyPages. <http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/W/Welcome.html>.

Kloczkowski, A. *et al*. "Combining the GOR V Algorithm With Evolutionary Information for Protein Secondary Structure Prediction From Amino Acid Sequence." Proteins: Structure, Function and Genetics 49 (2002): 154-166.

Korf, Ian *et al*. BLAST. San Francisco: O'Reilly and Associates, 2003.

Krane, Dan E and Michael L Raymer. Fundamental Concepts of Bioinformatics. San Francisco: Benjamin Cummings, 2003.

Kulharia, Mahesh *et al*. "Information Theory-Based Scoring Function for the Structure Based Prediction of Protein-Ligand Binding Affinity." Journal of Chemical Information and Modeling 48 (2008): 1990–1998.

Kuznetsov, Igor B and S Rackovsky. "On the Properties and Sequence Context of Structurally Ambivalent Fragments in Proteins." Protein Science 12 (2003): 2420-2433.

Lenaerts, Tom *et al*. "Quantifying information transfer by protein domains: Analysis of the Fyn SH2 domain structure." BMC Structural Biology 8.43 (2008).

Levinthal, Cyrus. "How to Fold Graciously." Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. Ed. J T P DeBrunner and E Munck. University of Illinois Press, 1969. 22-24.

Lim, V I. "Algorithms for Prediction of α-Helical and β-Structural Regions in Globular Proteins." Journal of Molecular Biology 88 (1974b): 873-894.

—. "Structual Principles of the Globular Organization of Protein Chains: A Stereochemical Theory of Globular Protein Secondary Structure." Journal of Molecular Biology 88 (1974a): 857-872.

MacKay, David J C. Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press, 2003.

Martin, L C *et al*. "Using information theory to search for co-evolving residues in proteins." Bioinformatics 21.22 (2005): 4116–4124.

Massachusetts Institute of Technology. "3.4 Primary through Quarternary Structure." Introductory Biology 7.01 Hypertextbook. 31 Jan 2011 <http://web.archive.org/web/20060411120350/web.mit.edu/esgbio/www/lm/proteins/structure/structure.html>.

Meiler, Jens and David Baker. "Coupled Prediction of Protein Secondary and Tertiary Structure." Proceedings of the National Academy of Science (PNAS) 100.21 (2003): 12105-12110.

Mezei, Mihaly. "Chameleon sequences in the PDB." Protein Engineering 11.6 (1998): 411-414.

Minor, Daniel L and Peter S Kim. "Context-dependent secondary structure formation of a designed protein sequence." Nature 380 (1996): 730-734.

Montgomerie, Scott *et al*. "Improving the accuracy of protein secondary structure prediction using structural alignment." BMC Bioinformatics 7.301 (2006).

Naderi-Manesh, H *et al*. "Prediction of Protein Surface Accessibility with Information Theory." Proteins: Structure, Function, and Genetics 42 (2001): 452–459.

National Institute of General Medical Sciences. "The Structures of Life." National Institute of General Medical Sciences. 30 Jan 2011 < http://www.nigms.nih.gov>.

National Institutes of Health. "Stem Cells: Scientific Progress and Future Research Directions 2001 appendix A ." National Institutes of Health. 30 Jan 2011 <http://stemcells.nih.gov/info/scireport/appendixA.asp>.

Nolting, Bengt. Protein Folding Kinetics: Biophysical Methods . Berlin: Springer-Verlag, 2006.

Pan, Xian-Ming. "Multiple Linear Regression for Protein Secondary Structure Prediction." PROTEINS: Structure, Function and Genetics 43 (2001): 256-259.

Pauling, Linus and Robert B Corey. "Atomic Coordinates and Structural Factors for Two Helical Configurations of Polypeptide Chains." Proceedings of the National Academy of Science (PNAS) 37 (1951b): 235-240.

—. "The Pleated Sheet, A Layer Configuration of Polypeptide Chains." Proceedings of the National Academy of Science (PNAS) (1951c): 251-256.

Pauling, Linus, Robert B Corey and H R Branson. "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain." Proceedings of the National Academy of Science (PNAS) 37 (1951a): 205-211.

Petersen, Thomas Nordahl *et al*. "Prediction of Protein Secondary Structure at 80% Accuracy." PROTEINS: Structure, Function and Genetics 41 (2000): 17-20.

Platt, John C. "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines." Microsoft Research Technical Report MSR-TR-98-14 (1998).

Pollastri, Gianluca and Aoife McLysaught. "Porter: a new, accurate server for protein secondary structure prediction." Bioinformatics Applications Note 21.8 (2005): 1719-1720.

Pollastri, Gianluca *et al*. "Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles." PROTEINS: Structure, Function and Genetics 47 (2002): 228-235.

Qin, Sanbo *et al*. "Predicting Protein Secondary Structure and Solvent Accessibility with an Improved Multiple Linear Regression Method." PROTEINS: Structure, Function and Bioinformatics 61 (2005): 473-480.

Racz, Attilla. Considerations on the kinetics of protein folding. Jan 2007 <http://www.chaperone.sote.hu/mainindex.html>.

Raymer, Michael L. Algorithms for Bioinformatics CS/Bio 471, Protein Slides. 2006. 30 Jan 2011 <http://birg.cs.wright.edu/cs471/>.

Research Collaboratory for Structural Bioinformatics (RCSB)Protein Data Bank. RCSB Protein Data Bank 2008 Annual Report. 2008. December 2009 <http://www.rcsb.org/pdbstatic/general_information/news_publications/annual_reports/annual_report_year_2008.pdf>.

Rost, B. 22 May 2000. < http://www.rostlab.org/eva/doc/measure_sec.html >.

Rost, Burkhard and Chris Sander. "Prediction of Protein Secondary Structure at Better than 70% Accuracy." Journal of Molecular Biology 232 (1993): 584-599.

Rost, Burkhard *et al*. "Redefining The Goals Of Protein Secondary Structure Prediction." Journal of Molecular Biology 235 (1994): 13-26.

Rost, Burkhard. "Rising Accuracy of Protein Secondary Structure Prediction." Protein Structure Determination, Analysis, and Modeling for Drug Discovery. Ed. D Chasman. New York: Dekker, 2003. 207-249.

Schiffer, Marianne and Edmundson Allen. "Use of Helical Wheels to Represent the Structures of Proteins and to Identify Segments with Helical Potential." Biophysical Journal 7 (1967): 121-135.

Shannon, C E. "A Mathematical Theory of Communication." The Bell Systems Technical Journal 27.July, October (1948): 379-423,623-656.

Shatsky, Maxim *et al*. "FlexProt: Alignment of Flexible Protein Structures without a Predefinition of Hinge Regions." Journal of Computational Biology 11.1 (2004): 83-106.

Shibuya, Tetsuo. "Fast Hinge Detection Algorithms for Flexible Protein Structures." IEEE/ACM Transactions on Computational Biology and Bioinformatics 7.2 (2010): 333-341.

Stryer, Lubert. Biochemistry. 4th Edition. New York: W H Freeman and Company, 1995.

Subair, S O A and Safaai Deris. "Protein Secondary Structure Reduction Methods Significantly Affect Prediction Accuracy." Proceeding AICCSA '06 Proceedings of the IEEE International Conference on Computer Systems and Applications. Washington DC: IEEE, 2006. 296-299.

Sudarsanam, Sucha. "Structural Diversity of Sequentially Identical Subsequences of Proteins: Identical Octapeptides Can have Different Conformations." Proteins: Structure, Function, and Genetics 30 (1998): 228-231.

Swanson, Rosemarie *et al*. "An Information Measure of the Quality of Protein Secondary Structure Prediction." Journal of Computational Biology 15.1 (2008): 65-79.

Tankano, Kazufumi *et al*. "Conformational Contagion in a Protein: Structural Properties of a Chameleon Sequence." Proteins 68 (2007): 617-625.

Ting, Kia Ming and Ian H Witten. "Stacking Bagged and Dagged Models." Proceedings of the 14th International Conference on Machine Learning. 1997. 367-375.

Wang, Long-Hui *et al*. "Predicting Protein Secondary Structure By A Support Vector Machine Based on a New Coding Scheme." Genome Informatics 15.2 (2004): 181-190.

Witten, Ian H and Eibe Frank. Data Mining. 2nd Edition. San Franciso: Morgan Kaufmann, 2005.

Wood, Matthew J and Jonathan D Hirst. "Protein Secondary Structure Prediction with Dihedral Angles." PROTEINS: Structure, Function, and Bioinformatics 59 (2005): 476-481.

Wriggers, Willy and Klaus Schulten. "Protein Domain Movements: Detection of Rigid Domains and Visualization of Hinges in Comparisons Atomic Coordinates." PROTEINS: Structure, Function, and Genetics 29 (1997): 1-14.

Xiang, Zhexin. "Advances in Homology Protein Structure Modeling." Current Peptide Science 7.3 (2006): 217-227.

Yi, Tau-Mu and Eric S Lander. "Protein Secondary Structure Prediction using Nearest-neighbor Methods." Journal of Molecular Biology 232 (1993): 1117-1129.

Zelma, Adam *et al*. "A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment." PROTEINS:Structure, Function, and Genetics 34 (1999): 220-223.

Zhou, Xianghong *et al*. "An Analysis of the Helix-to-Strand Transistion Between Peptides with Identical Sequence." Proteins: Structure, Function, and Genetics 41 (2000): 248-256.

APPENDIX A
SECONDARY STRUCTURE PREDICTION

ACCURACY MEASURES

ACCURACY MEASURES

Before a meaningful comparison of secondary structure methods can be performed, one or more measures of quality must be defined. Rost has identified several such metrics from the literature. These include: 1) the prediction accuracy matrix; 2) the three state accuracy measure (Q_3); 3) per-state percentage; 4) Segment Overlap measure (SOV); 5) Mathew's Correlation Coefficient; and 6) reliability index[Rost, 2000].

A.1 Prediction Accuracy Matrix

The prediction accuracy matrix is also known as the contingency or "confusion" matrix.

		Predicted			
		Helix	Extended	Coil	Total
Actual	Helix	N_{hh}	N_{he}	N_{hc}	$T_{actualh}$
	Extended	N_{eh}	N_{ee}	N_{ec}	$T_{actuale}$
	Coil	N_{ch}	N_{ce}	N_{cc}	$T_{actualc}$
	Total	T_{predh}	T_{prede}	T_{predc}	T_{total}

Table 30 - Contingency Matrix

The prediction accuracy matrix compares the number (N_i) of actual (observed) residues in a particular conformation with the predicted secondary structure. A number of measures can be computed directly from the information in this matrix.

A.2 Three state accuracy (Q_3)

Three state accuracy (Q_3) is the most commonly used measure for secondary structure prediction. Nearly all papers in the literature report at least this measure. Expressed as a percentage, it is computed by dividing the number of correctly predicted residues by the total number of residues. Using the data from table 25 above:

$$Q_3 = 100 * (N_{hh} + N_{ee} + N_{cc})/ T_{tot}$$

A.3 Per State Percentage (PSP)

Often the per state percentage is also reported. It is computed by dividing the number correctly predicted in each state (H,E,C) by the number of residues in each state (H,E,C). Using the data from table 25:

$$PSPh = 100 * (N_{hh} / T_{ah})$$

$$PSPe = 100 * (N_{ee} / T_{ae})$$

$$PSPc = 100 * (N_{cc} / T_{ac})$$

A.4 Segment Overlap (SOV)

Another popular measure of prediction accuracy is the segment overlap measure or SOV. Rost, Sanders, and Schneider developed the first version of the SOV in 1994 [Rost *et al.,* 1994]. This was later improved by Zelma *et al*.,1999. Both versions are aimed at correcting problems presented by Q_3. These include: the type and position of segments; the natural variation of segment boundaries among homologous proteins; and ambiguity in the position of segment ends due to differences in secondary structure classification [Zelma *et al.*, 1999. SOV equations from Rost, 2000].

Per-stage segment overlap:

$$SOV_i = \frac{1}{N_i} \sum_{s_i} \frac{MINOV(S1;S2) + DELTA(S1;S2)}{MAXOV(S1;S2)}$$

with the following definitions:

S1 and S2 are the observed and predicted secondary structure segments (in state i, which can be either H, E or C)

LEN(S1) is the number of residues in the segments S1

MINOV(S1;S2) is the length of actual overlap of S1 and S2, i.e. the extent for which both segments have residues in state i, for example H

MAXOV(S1;S2) is the length of the total extent for which either of the segments S1 or S2 has a residue in state i

DELTA(S1;S2) is the integer value defined as being equal to the following

$$DELTA(S1;S2) = \min \left\langle \begin{array}{l} MAXOV(S1;S2) - MINOV(S1;S2) \\ MINOV(S1;S2) \\ INT(0.5 \cdot LEN(S1)) \\ INT(0.5 \cdot LEN(S2)) \end{array} \right\rangle$$

THE SUM (Σ) is taken over S, all the pairs of segments {S1;S2}, where S1 and S2 have at least one residue in state i in common

N(i) is the number of residues in state i defined as follows:

$$N_i = \sum_{S(i)} LEN(S1) + \sum_{S'(i)} LEN(S1)$$

The two sums are taken over S and S':

S(i) is the number of all the pairs of segments {S1;S2}, where S1 and S2 have at least one residue in state i in common

S'(i) is the number of segments S1 that do not produce any segment pair

Segment OVerlap quantity measure for all three states: where the normalization value N is a sum of N(i) over all three conformational states (i = HELIX, STRAND, COIL)

A.5 Matthews correlation coefficient

Occasionally, Matthews' correlation coefficient (MCC) is reported. The Matthews correlation coefficient can be computed directly from the contingency table using the following formula:

$$MCC = (T_p * T_n - F_p * F_n) / [(T_p + T_n)(T_p + F_n)(T_n + F_n)(T_n + F_n)]^{1/2}$$

Where:

T_p = the number of true positives = N_{ii} for a given row i
T_n = the number of true negatives = ΣN_{jk} where $j \neq i$ and $k \neq i$ for a given row i
F_p = the number of false positives = ΣN_{ik} where $k \neq i$ for a given row i
F_n = the number of false negatives = ΣN_{ki} where $k \neq i$ for a given column i

As one can see there will be three MCCs computed, one each for helix, extended and coil. In the few cases where the denominator = 0, the numerator will necessarily = 0 and MCC is defined as 0. In all other cases, MCC will be in the range -1 to 1 inclusive with 1 being a perfect classifier, -1 being a classifier that is always wrong and 0 being random, much like a Pearson's R.

A.6 Reliability Index

Another measure which is often reported is the reliability of a prediction. This is the difference in probability between the most likely state (H,E,C) and the next most likely state (HEC) for an individual residue. It is usually reported as the first significant digit of the difference in probabilities. Hence, it runs from 0-9. The reliability index is calculated by residue and is often used as a measure of confidence in individual predictions. It is often used to bin predictions. Statements like, "The Q_3 for predictions on residues with a reliability index greater than 7 is 83%." are typical.

The most popular of these measures is the Q_3 measure. The SOV and Matthew's correlation measures appear occasionally. Some researchers also report the percent of each segment of a particular length which has been correctly predicted.

APPENDIX B

SECONDARY STRUCTURE PREDICTION

COMPARING METHODS

COMPARING METHODS

Given the wide variety of approaches that have been taken to the problem of secondary structure prediction over the years, comparison of similar methods can be problematic. In order to organize and categorize the wide array of approaches appearing in the literature, I have identified here several axes of variation with which can be used to compare methods in the area of secondary structure prediction.

1. What is the model or method?
2. What data is used?
3. What 8 to 3 reduction is used?
4. What is the unit of analysis?
5. What transformations are conducted?
6. How is the model/method validated?
7. How transparent is the model?
8. How accurate are the predictions?

Table 31 - Key Questions

B.1 What is the model or method?

There are literally hundreds of papers on secondary structure prediction. They can be divided into three broad classes based on the underlying methods used to accomplish the prediction. These are 1) physico-chemical; 2) homology based and 3) ensemble methods. Most modern methods will combine elements of more than one of these categories.

B.1.1 Physico-chemical

The physico-chemical models are based on the physical properties of the amino acids such as size, hydrophobicity, charge, position and aromaticity. Examples include: Lim, helical wheels and molecular dynamics models.

B.1.2 Homology based

Homology based methods operate on two ideas: that the primary structure determines the secondary and tertiary structures; and that tertiary structure is conserved through evolution. This means that minor changes in the primary sequence will generally not be reflected in the tertiary structure. In fact, it has been shown that:

> "When the sequence identity is above 40%, the alignment is straight forward, there are not many gaps, and 90% of main-chain atoms could be measured with a RSMD (root-mean-square distance) of about 1 Å. …When the sequence identity is about 30-40%, obtaining correct alignment becomes difficult, where insertions and deletions are frequent. For sequence similarity in this range, 80% of main-chain atoms can be predicted to RMSD 3.5 Å, while the rest of residues are modeled with large errors, especially in insertion and deletion regions. …When the sequence similarity is below 30%, the main problem becomes the identification of the homologue structures, and alignment becomes much more difficult." [Xiang, 2006 p 217]

As the above quote shows, above 30% similarity the tertiary structure can be estimated quite well. Below this level, predicting secondary structure is a very useful intermediate step. Two types of homology based models are statistical methods and pattern recognition methods.

B.1.2.1 Statistical methods

The statistical methods include: frequency based models such as Chou-Fasman and Bayesian models; information based models like GOR and several datamining tools; and linear models including linear discriminant analysis and regression.

B.1.2.2 Pattern recognition

Pattern recognition is a broad area which includes artificial neural networks, nearest neighbor methods, and hidden Markov models. Many of the most successful individual classifiers are pattern recognition methods.

B.1.3 Ensemble methods

Several of the best overall classifiers explicitly combine the results of other methods into a 'meta' prediction. Examples include JPRED and cascaded classifiers. Various voting schemes are used to take advantage of the strengths and weaknesses of different models.

B.2 What data is used?

The source, format, degree of homology/redundancy, accuracy and completeness of the data are all important considerations when comparing prediction methods. The history of secondary structure prediction is replete with bold claims which turned out to not generalize due to limited data.

B.3 What 8 to 3 reduction is used?

As shown above**,** several reduction methods can be used to convert DSSP data (or STRIDE or DEFINE data) to helix, extended and coil structures. One may select a reduction method for theoretical reasons or to focus on a particular structure (predicting strands for example). Since the choice can affect the measured accuracy of the prediction by up to three percent, it is important that one compares predictions with similar reduction methods.

B.4 What is the unit of analysis?

The basic unit of analysis used by a prediction method in its internal calculations is often critical to its success. Rost has defined three generations of secondary structure predictions based on the unit of analysis [Rost 2003]. The first generation is based on single reside statistics; the second on segment statistics; and the third on multiple alignment data. Examples of first generation classifiers include Chou-Fasman, GOR I, and Lim. Second generation classifiers include GORIII and COMBINE. Third generation methods include PHD, PSIPRED, and JPRED2. With each generation, the information required increases by a rough order of magnitude. Accuracy also increases. The first generation classifiers are generally between 50 and 60% accurate; the second generation between 60% and 70% and third generation above 70%.

B.5 What transformations are conducted?

Directly related to the methods question is the number and type of data transformations. A common transformation technique is windowing. Others are to sum, average, difference or smooth some property over a local region. Frequencies are often normalized. Ideally, a transformation will highlight a relationship while suppressing or eliminating noise in the data. What transformations are conducted (or not) under what conditions can greatly affect the result.

B.6 How is the model/method validated?

Most methods for validating secondary structure prediction techniques can be classified into one of two general categories: holdout testing and cross-validation. The first method is to independently validate the method using an unrelated data set in which no member

of the test set has been used in developing or training the method. This is sometimes difficult, given that data representing demonstrably independent, non redundant, non-homologous proteins is often in short supply. If one uses a significant portion as an independent test set, it is not available for model development.

In order to address this, most researchers use cross validation. N-way cross validation divides the data into N groups. Each group is held out and the model is developed/ trained on the remaining N-1 groups. This is repeated until all groups have participated as the test group. The results from the N tests are then averaged to get a validated result. Seven and ten are the Ns most often used in the literature to cross validate. An extreme variant of this approach is called 'jack knife' or 'leave one out' cross validation. Here the number of groups (N) is equal to the number of individual cases in the data set. Hence, one removes a single test case and builds the model using the remaining cases. This is repeated until all cases have been tested. The results are then averaged. This can be very time consuming if the number of cases is large.

B.7 How transparent is the model?

There are two types of transparency which are important in secondary structure prediction. 1) Are the inner workings of the model easily understood? 2) Does the model give insights into the physical phenomenon it attempts to depict? These are clearly related. There is often a tradeoff between accuracy and transparency. Phyiso–chemical methods often do very well on transparency but may be computationally intensive (molecular dynamics) or limited in their accuracy (helical wheels). Neural

networks, on the other hand are often highly accurate (>70%) but may appear as 'black boxes' offering very little visibility into the physics of the problem.

B.8 How accurate are the predictions?

Of all the accuracy measures given above Q_3 and SOV are the two measures usually used to compare prediction results in the literature. However**,** prior to comparing the accuracy claims of any two methods, one should ensure that they are running against the same dataset, using the same reduction algorithm, and the same validation method. This is sometimes not done, resulting in apples and oranges comparisons.

APPENDIX C

SECONDARY STRUCTURE PREDICTION

RESEARCH

RESEARCH

C.1 Early Explorations

C.1.1 Longest Matching String

One of the key assumptions behind homology based secondary structure prediction is that proteins of similar sequence will adopt similar structures. Therefore, it may be possible to determine the structure of a novel protein by comparison with existing proteins. Taking this idea to its logical endpoint, a straightforward technique for predicting secondary structure would be to identify a region of sequence, then find similar sequences in proteins of known structure, and predict the same secondary structure observed in these proteins for the given protein. As a preliminary investigation into the effectiveness of this sort of direct approach, I devised a longest matching string algorithm for direct-comparison-based secondary structure prediction. This exploration was based on the following assumptions:

1) there are a limited number of strings in nature;
2) the longer the string, the more unique the secondary structure; and
3) the relative frequency of each structure in each position would provide a good classifier.

Given that there are 20 amino acids, if they were put together randomly, there are 20^n potential sequences of length n. To find out how many combinations actually appeared in the PDB, a program was written to count the number of different sequences of length n from 1 to 28. (PDB, Jan 05)

Figure 30 shows the results of that program. The number of different secondary structures grows at approximately 20 to the nth until it reaches n = 6. It stabilizes at just over 2 million.

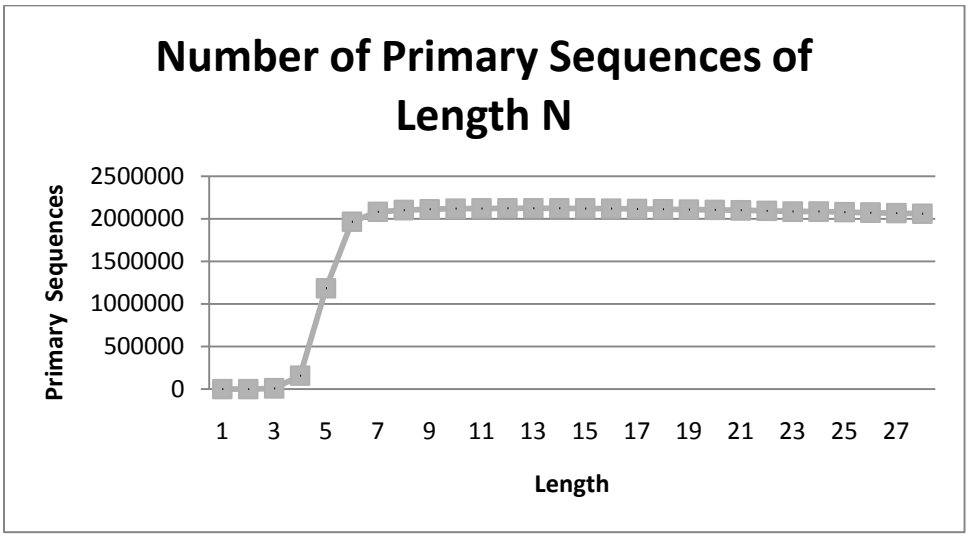


Figure 30 - Number of Different Primary Structures of Length N (Jan 05)

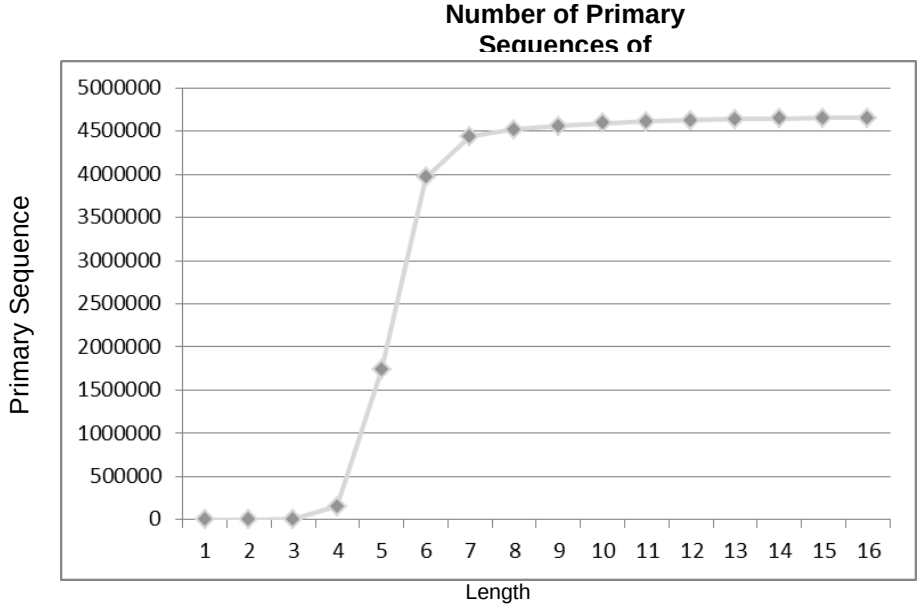


Figure 31 - Number of Primary Sequences of Length N (Mar 09)

The PDB has experienced very rapid growth in recent years (Tables 4 and 5). The count was retaken four years later and the result appears in Table 31. Here too, the number of different sequences levels off when n = 6.

This is an important finding. If the number of different sequences which appear in nature were anywhere near 20 to nth for a moderately sized n this approach would quickly fail due to computational space and time limitations. Instead, this initial investigation showed that the number of different sequences grows exponentially until a near constant is reached.

C.1.2 Results

The longest matching string approach was tried on a few proteins with disappointing results. The predictions had a Q_3 of approximately 55%. A possible explanation for the underperformance of this method includes:

1) The PDB is highly redundant. Therefore, the simple counts of secondary structure associated with each sequence are highly suspect.
2) Minor differences in sequences were not addressed (Gaps, insertions, deletions, etc.).
3) Problems of identifying and controlling for homology were not addressed at all.
4) When the longest match program was run against a properly constructed database (CB513) the longest match identified was often only 5 in length, demonstrating that the approach, as implemented, had very limited usefulness.

C.1.3 Additive Windows

Using the lessons learned from the longest matching string algorithm, I formulated a second algorithm based on the idea of using information from multiple sliding windows together to predict secondary structure. In this analysis the database used was not the PDB, but rather CB513. This eliminated the problems with redundancy and excessive homology.

Each amino acid in a sequence which is at least one window length (*l*) away from the ends appears in *l* windows. For example, with a window length of three, all of the amino acids with the exception of the first two and the last two in a given sequence participate in three windows. When the window length is five, all but the first and last four amino acids appear in five windows.

Matching the windowed amino acids with a set of known structures as in the longest matching algorithm, *l* predictions are created for each position in the sequence based on the most frequent structure at that position in the windowed known data. There are five predictions for each amino acid using a length of five and four predictions with a window of four etc. As a result, if all windows were found in the known database there would be a maximum of Σ i where i = 1,2…5 or 15 separate predictions for each amino acid. For those near the ends of the sequence the maximum clearly would be less with the terminal amino acids having five predictions.

The results for this method on the few proteins tested were also poor. (Q_3 of 54%.) However, two things were learned. 1) There were few sequences of length five which

had matches in the data base, making the five-window of limited value. 2) The windows of length three dominated all others. If any of the other windows predicted correctly the three-window also predicted correctly. If the window of length three was wrong so were the others. As a result of this exploration, this method was also abandoned.

C.2 Candidate Predictor

Before formulating a new method I again reviewed the current literature. Based on this review the following observations can be made:

1) The most successful methods are neural networks or combinations of neural networks;
2) The use of multiple alignments, particularly Position Specific Scoring Matrices (PSSM), are key to the most successful methods.
3) Windows are 9 to 17 long.
4) Most researchers use a customized data base with CB513 as one of the more popular.

The next attempt incorporates each of these ideas. After a review of readily available neural networks / pattern recognition packages, the Waikato Environment for Knowledge Analysis (WEKA) was selected. It is powerful, has an excellent user interface, and is in relatively widespread use. It also implements many data analysis and pattern recognition techniques making it easy to experiment with different alternatives. The National Center of Biological Information (NCBI)'s BLAST was used to generate the position specific

scoring matrix and the main analysis was accomplished using a window of length 13. Figure 32 below depicts the new method.

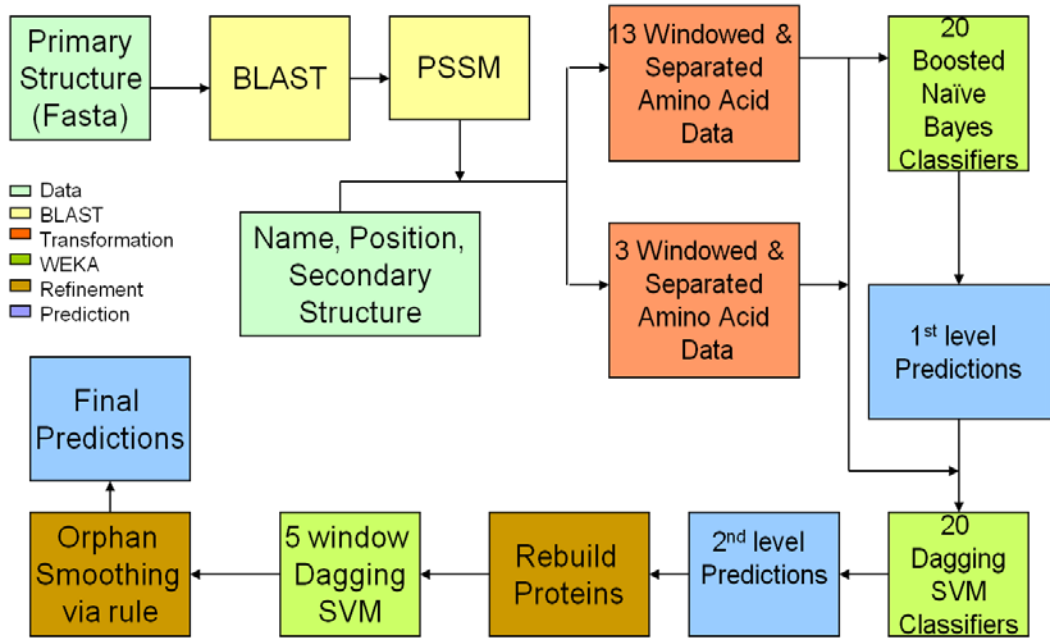


Figure 32 - Candidate Prediction Method

C.2.1 Data

The process starts with the Cuff and Barton's 513 database. The CB513 database was carefully constructed to address redundancy and homology issues, eliminating one of the difficulties in the earlier attempts. In addition, a number of papers have used this database and so provide a reasonable basis of comparison. Finally, CB406 provides a pre-established validation or test set if one is required. These data sets are available at http://www.compbio.dundee.ac.uk/jpred_v2/data/

C.2.2 BLAST

The Basic Local Alignment Search Tool (BLAST) is an automated tool which attempts to answer the question "How similar is this sequence to other sequences in my library?"

In 1970 Needleman and Wunsch devised a dynamic programming approach to aligning two protein (or DNA) sequences. Aligned sequences can then be assessed for similarity. In 1981 Smith and Waterman modified the whole-sequence (global) method of Needleman and Wunsch, devising a novel local alignment method that can find and align the best subsequence between two sequences. By concatenating all known sequences together into a single database, the local alignment method can be used to search a database for the closest match for a new sequence. However, the run time is intractable for such a large sequence search.

The BLAST algorithm, devised by Altschul *et al*. in 1990, takes several shortcuts in searching for close sequence matches. As a result, BLAST is not guaranteed to find all close sequence matches of a given query sequence, but the run time is significantly lower than a complete dynamic programming approach.

The BLAST algorithm breaks the alignments into short sequences of high frequency "words" and searches the library for matches. Once found, BLAST then tries to build the alignment in both directions until the end is found or the similarity score falls below some threshold. While not as sensitive as Needleman-Wunsch or Smith-Waterman, the BLAST algorithm is computationally efficient enough to make large scale multiple alignments a practical reality. It is not restricted to amino acids in a protein, but is often used to study sequences of nucleotides in DNA. BLAST is available from the National

Center of Biological Information web site http://www.ncbi.nlm.nih.gov/. [Krane and Raymer 2003; Korf *et al.* 2003]

C.2.3 PSSM

The primary structure of each protein is searched against a sequence database using PSI-BLAST to generate the position specific scoring matrix. The PSSM data is combined with the protein name, amino acid position and secondary structure information from the CB513 files.

C.2.4 Windows

The data is then windowed twice, once using a 13 amino acid window and a second time using a three amino acid window. The thirteen amino acid window was selected because some researchers [Rost and Sanders 1993] have found that a window size of 13 is the most effective for secondary structure prediction. However, a window of size 13 leaves six amino acids on each end of the input sequence which cannot be predicted due to insufficient data. Some researchers resolve this by tagging the affected amino acids with a special flag. In this effort the ends were predicted using a three window analysis. Combining the results of the 13 and 3 windowed analyses leaves only the very first and last amino acids as not computable. Fortunately, the first and last amino acids of most proteins are overwhelmingly coil (95%). This allows the secondary structure of the entire protein to be predicted.

C.2.5 Twenty classifiers

The data is then separated into twenty files based on the central amino acid, one for each amino acid. Hence, all of the alanines are in one file all the valines are in another. The origin of each central amino acid (protein and position) is maintained. These partitioned data are then used to train classifiers devoted only to their respective amino acid, resulting in twenty separate classifiers per level. This is unique among current prediction methods.

C.2.6 WEKA

The Waikato Environment for Knowledge Analysis machine learning workbench (WEKA) is a software package developed by Witten and Frank at the University of Waikato, New Zealand, to conduct data mining. As of this writing, it implements nearly fifty classifier algorithms, two dozen meta-leaning algorithms, five clustering algorithms, and three association rule learners. In addition, it has numerous built-in tools for data visualization, exploration and analysis. It is free software available at http://www.cs.waikato.ac.nz/ml/weka/. I chose to use its implementation of a boosted naïve Bayes classifier and a support vector machine[Witten and Frank 2005].

C.2.6.1 Boosted Naïve Bayes Classifiers

The data for each amino acid is then fed into boosted naïve Bayes classifier. A Bayes classifier is one based on Bayes' Rule namely,

$$P(x \mid y) = \frac{P(x) * P(y \mid x)}{P(y)}$$

A naïve Bayes classifier is one which assumes that the values of each of the variables are independent of each other. One method used to improve the generality of a classifier is called bagging. Bagging takes several samples from a data set and trains a different instance of the classifier on each subset. The results are then combined to form a global prediction. Boosting is a variant of bagging which builds a succession of classifiers giving higher weight to those classifiers which are more accurate. It also attempts to select a greater percentage of misclassified samples to enable additional learning. [Witten and Frank 2005].

C.2.6.2 Dagged Sequential Minimal Optimization Support Vector Machine

A dagged sequential minimal optimization support vector machine is used twice in the candidate design (Figure 32).

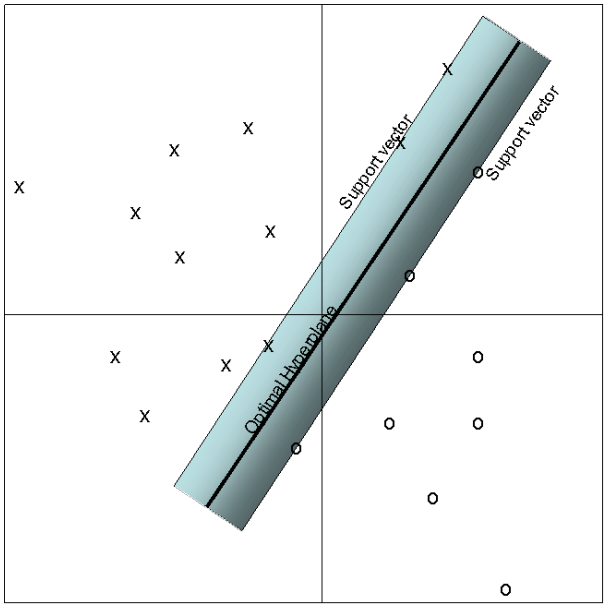


Figure 33 - Idealized Support Vector Machine
adapted from Haykin 2005, p.320

A support vector machine is a machine learning technique which classifies instances based on identifying the hyperplane which separates the classes to the greatest extent possible. It does this by computing the support vectors with the largest total margin from the optimal hyperplane. An idealized support vector machine using linearly separable is shown in Figure 33.

A support vector machine initially maps the data into a higher dimension feature space and then constructs the hyperplane that best classifies the data. This mapping may be done by a number of non-linear functions. This allows many non-linear problems to be resolved. Haykin lists three commonly used functions: polynominal, radial basis, and two layer perceptron [Haykin 2005, p. 333].

The identification of the optimal hyperplane is usually done using quadratic programming(QP). The sequential minimal optimization (SMO) algorithm accomplishes this by breaking the problem into a series of smaller QP problems which can be solved analytically and therefore more quickly with less computational space [Platt 1998].

Finally, dagging is the same as bagging except that instead of creating samples which may overlap (bagging) all of the samples used in dagging are disjoint [Ting and Witten 1997].

C.2.7 Three levels of predictions

The first level predictions from the boosted naïve Bayes classifiers are added to the data and are fed into the second level classifiers. Each of the second level classifiers is a

dagged sequential minimal optimization support vector machine**.** The results of the second level classifiers are then added to the data and fed into another support vector machine. Where the first and second level classifiers use a window of 13 amino acids, the third level classifier uses a window of length 5.

C.2.8 Rebuild Proteins

The next step is to rebuild the proteins. When the proteins were separated into the twenty different amino acid databases their original protein and position were attached to each instance. These are now used to rebuild the proteins.

C.2.9 Orphan Smoothing Rule

Unfortunately, this splitting out and recombining of amino acids, makes this technique subject to some errors that other methods do not experience. One of these errors is what I term an "orphan". An orphan is where a singleton structure is found between two structures of a different type. For example, given ...HHHEHH…, it is clear that the extended structure in the middle of a helix is an error. To address this, a program was written to find and convert orphans to the same structure as their neighbors.

Whether one searches for orphans from the N or the C terminus can make a difference. Most orphans are directionless, that is it does not matter whether one comes to them from the front or the back. For some however, it is important. As one can see from Figure 34 below, if one comes from the front the first H will be flipped to an E. If one comes from the back the second E will flipped.

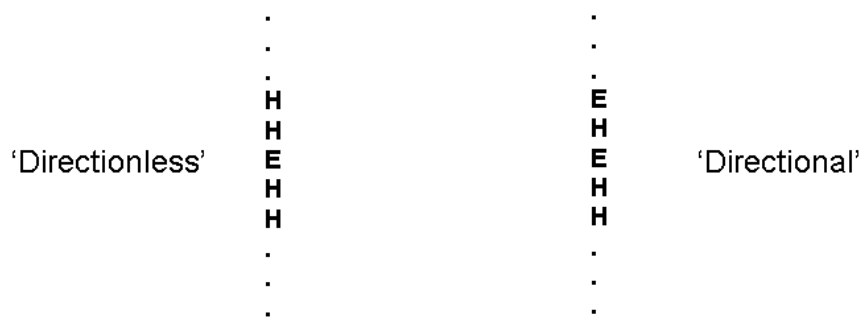


Figure 34 - Orphan Types

This problem has no clean resolution as of now. An arbitrary decision was made to search from front to back. Comparing the results to a backward to front search showed that this was marginally better than the reverse, but given a different data set to predict it clearly could be otherwise.

C.2.10 Results

A ten-fold cross validation was used on the each of the twenty amino acid classifiers. The unit of validation was a protein not the amino acid. Fifty-one proteins were placed in the test set for seven of the test sets and fifty two were placed in three test sets. Therefore, each amino acid appeared exactly once in a test set. The test results were gathered and proteins were rebuilt. The orphan smoothing was then applied.

The Q_3 for this effort was 75.3% on a residue basis.

APPENDIX D

EARLY EXPLORATION ALGORITHMS

EARLY EXPLORATION ALGORITHMS

D.1 Longest Matching String Algorithm

The longest matching string algorithm operates as follows:

Given:

X = Protein sequence of unknown structure

K = Set of sequences with known structure

s_i = ith string of amino acids

T_i = ith secondary structure

$S_X = \{s_i \mid s_i \in \text{of } X\}$ i = 1,2,…n

$S_K = \{(s_i,T_j) \mid s_i \in K \text{ and } s_i \rightarrow T_j\}$ i = 1,2,…n and j = 1,2,…m

$M = S_X \text{ -> } S_K$ = Set of matching strings

M_L = si | si $\in$M and length (si) $\geq$ length (sj) $\forall$ sj $\in$M; j = 1,2,…n

1) Find M_L in X

a) define a sliding window of length L where L = length of X

b) check if a match for sliding window occurs in K

i) if yes, report as M_L

ii) if no, reduce L by 1 and repeat until found or L = 0

2) Partition X

X necessarily can be partitioned into three parts.

$X = X_F, M_L, X_B$

Where

X_F = sequence prior to M_L in X

X_B = sequence following M_L in X

3) Build M_K

Find all instances of M_L in S_K

$M_K = \{Ti \mid (M_L, T_i) \in \text{of } S_K\}$

4) Predict T_L

$(M_L, T_L) = (M_L, T_i)$ where T_i max(freq(M_K))

5) Repeat steps 1-4 with XF and XB until structures are predicted for all amino acids.

The first exploration was to analyze a protein's primary structure and identify the longest string which matched a string within a database of known structures using the algorithm outlined above. Two observations should be noted: L in step one will never equal zero since all proteins are assumed to have only the twenty common amino acids; and the prediction TL in step four can be computed on either a string or amino acid/position basis. The latter was used in this investigation.

D.2 Additive Windows Algorithm

X = Protein sequence of unknown structure

K = Set of sequences with known structure

S_{xi} = Set of strings within X that are of length i

S_{ki} = Set of strings within K of length i

Tki = Set of structures within K associated with a string of length i

Wij = string defined by window of length i in position j

Fhpij = Frequency of helix in position p in a string defined by window of length i in position j found in S_{ki}

Fepij = Frequency of extended in position p in a string defined by window of length i in position j found in Ski

Fcpij = Frequency of coil in position p in a string defined by window of length i in position j found in Ski

Rpij = Prediction of structure for amino acid in position p using window of length i in position j

Rpi = Prediction of structure of amino acid in position p combining the results of all windows of length i

Rp = Prediction of structure of amino acid in position p combining the results of all windows of all lengths (1…5)

1. Find Rpij

Using a sliding window of length i in position j, find all instances of the window in Ski. Add one to the count for whichever structure is most frequent in position using this window and this position.

Given a sequence AVGTE…

The glycine in position 3 would participate in three windows of length 3 (AVG,VGT,GTE). Each of these is compared to the set of known structures and the frequency of each structure associated with G (Fhpij, Fepij, and Fcpij) in each of the strings is returned, with the most frequent labeled as the prediction.Rpij

2. Combine the predictions for windows of length i

Rpi = max(Fhpi, Fepi, Fcpi)

Fhpi $= \Sigma$ Fhpij for j = 1…i

Fepi $= \Sigma$ Fepij for j = 1…i

Fcpi $= \Sigma$ Fcpij for j = 1…i

3. Combine the predictions for all windows

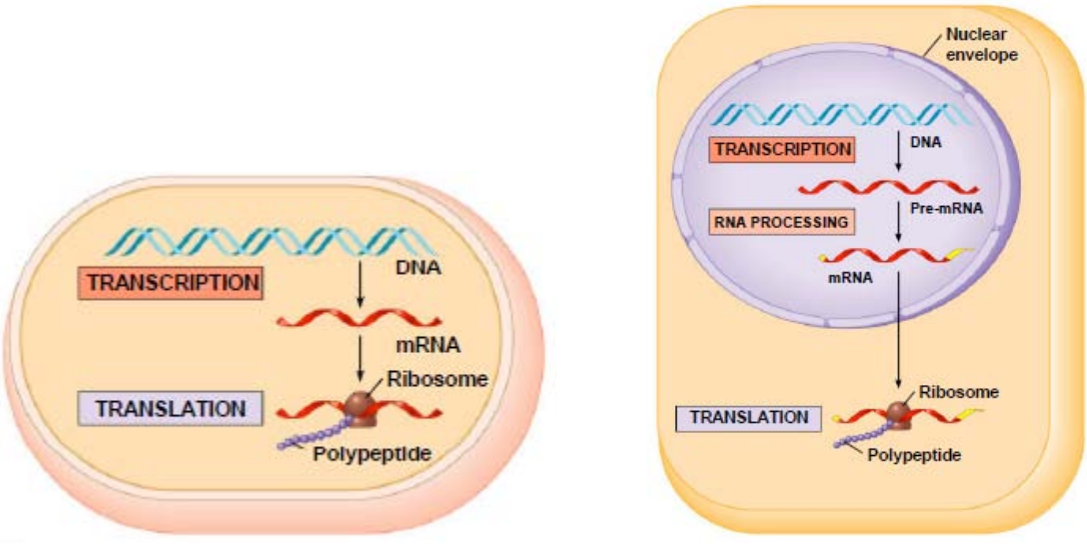
Rp = Σ Rpi for i = 1,2…5

APPENDIX E

PROTEIN SYNTHESIS

PROTEIN SYNTHESIS

In order to model protein structure at any level, it is important to have a basic understanding of the process through which proteins are synthesized. In most living cells, information necessary for protein synthesis is stored in molecular form through linear or circular chains of deoxyribonucleic acid (DNA). The constituent parts of this polymer are nucleic acids. The four most common, which form the alphabet of the genetic language, are adenine (A), cytosine (C), guanine (G), and thymine (T) arranged in a double helix. In fact, DNA is a recipe for making proteins. A protein is made from DNA in three key steps, transcription, translation and assembly.



a) Prokaryotic cell b) Eukaryotic cell

Figure 35 - Transcription and Translation
Campbell *et al.* 1999,p. 297

E.1 Transcription

The DNA is unwound used as a template to create a copy of the genetic information in messenger ribonucleic acid or mRNA. When this is done the thymine is transcribed as uracil(U). For those organisms with nuclei in their cells, termed eukaryotes, transcription occurs in the nucleus of the cell. Following the initial transcription the mRNA goes through some additional preparation and then exits the nucleus. Translation then occurs in the cytoplasm (Figure 35) [Campbell *et al.* 1999, p 297].

In organisms without a nucleus in the cells, prokaryotes, the both transcription and translation occur within the cell and there is no post processing to prepare the mRNA for the cytoplasm.

E.2 Translation

The messenger RNA is read by a ribosome. The “message” is encoded into three letter words called codons. Each three letter codon corresponds to one of the twenty naturally occurring amino acids. As the ribosome moves down the mRNA it translates the message into a list of amino acids required to build each protein. These messages all begin with the start codon (AUG) and continue until one of three stop codons is reached (UAA, UAG, and UGA).

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met, Start	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Table 32 - Standard Genetic Code

E.3 Assembly

Another form of RNA, called transfer RNA (tRNA) is found in the cytoplasm of eukaryotes and brings a specific type of amino acid to the ribosome based on the three letter code. As the ribosome reads the messenger RNA, it assembles the polypeptide chain, attaching the required amino acid in the order specified by the codons. In prokaryotes, transcription and translation are more closely linked. In bacterial ribosomes begin translation while transcription is occurring [Campbell, 1999, 296-297]. The ribosome assembles the primary structure of the protein. The next step in protein synthesis is the folding up of the linear polypeptide chain into the three dimensional native conformation for each protein. The native conformation is the end state for protein synthesis.

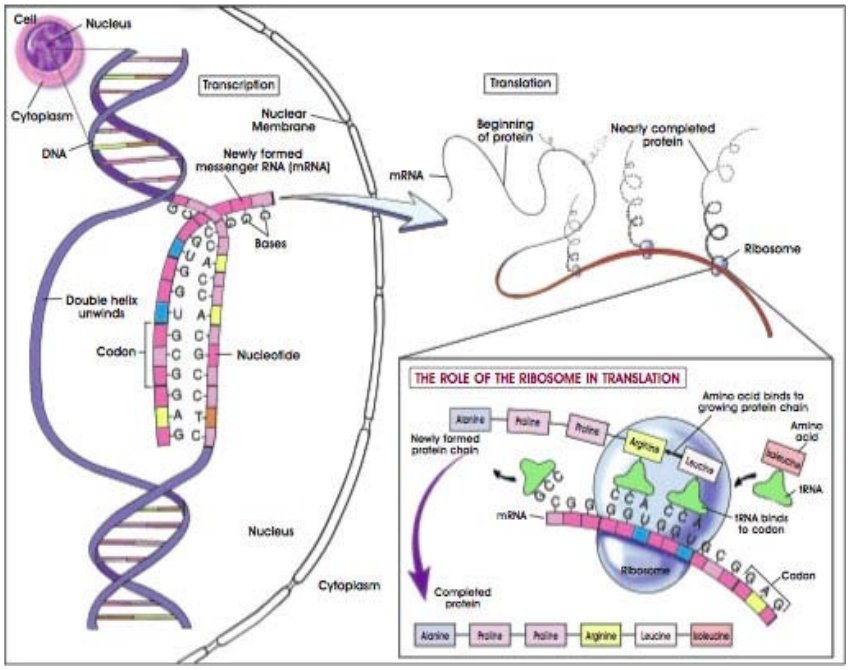


Figure 36 - Eukaryotic Protein Synthesis
National Institutes of Health, 2001

APPENDIX F

CATH DOMAINS

Table 33 - CATH Database - Number of Domains by Architecture
Version 3.3 Sept 29, 2010
http://www.cathdb.info/

1.0 Mainly Alpha

1.10	Orthogonal Bundle	19325
1.20	Up-down Bundle	6562
1.25	Alpha Horseshoe	576
1.40	Alpha solenoid	6
1.50	Alpha/alpha barrel	411

2.0 Mainly Beta

2.10	Ribbon	1306
2.20	Single Sheet	465
2.30	Roll	2771
2.40	Beta barrel	9805
2.50	Clam	17
2.60	Sandwich	15741
2.70	Distorted Sandwich	917
2.80	Trefoil	456
2.90	Orthogonal Prism	47
2.100	Aligned Prism	114
2.102	3-layer Sandwich	134
2.105	3 Propellor	1
2.110	4 Propellor	22
2.115	5 Propellor	46
2.120	6 Propellor	337
2.130	7 Propellor	225
2.140	8 Propellor	253
2.150	2 Solenoid	21
2.160	3 Solenoid	307
2.170	Beta Complex	484

3.0 Mixed alpha-beta

3.10	Roll	5662
3.15	Super Roll	5
3.20	Alpha-Beta Barrel	5544
3.30	2-Layer Sandwich	17965
3.40	3-Layer(aba) Sand.	26500
3.45	3-Layer(aab) Sandwich	0
3.50	3-Layer(bba) Sandwich	1276
3.55	3-Layer(bab) Sandwich	19
3.60	4-Layer Sandwich	1433
3.65	Alpha-beta prism	170
3.70	Box	53
3.75	5-stranded Propellor	73
3.80	Alpha-Beta Horseshoe	152
3.90	Alpha-Beta Complex	7487
3.100	RibosomalProtein L15; Chain K; domain2	116

4.0 Few Secondary Structures

4.1	Irregular	1883

Total 128688